



**ΣΧΟΛΗ : ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑ ΚΑΙ ΨΗΦΙΑΚΕΣ
ΚΑΙΝΟΤΟΜΙΕΣ**

ΤΙΤΛΟΣ ΔΙΠΛΩΜΑΤΙΚΗΣ

**«Ανίχνευση Κυβερνοεπιθέσεων με Τεχνικές Μηχανικής
Μάθησης: Σύγκριση και Ερμηνεία Αλγόριθμων
Ανίχνευσης Ανωμαλιών σε Δεδομένα Δικτύου»**

ΑΘΑΝΑΣΙΟΣ ΣΚΑΜΑΓΚΑΣ

ΜΙΧΑΗΛ ΓΕΩΡΓΙΑΔΗΣ

06/2026



**ΣΧΟΛΗ : ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΨΗΦΙΑΚΕΣ
ΚΑΙΝΟΤΟΜΙΕΣ**

ΤΙΤΛΟΣ ΔΙΠΛΩΜΑΤΙΚΗΣ

**«Ανίχνευση Κυβερνοεπιθέσεων με Τεχνικές Μηχανικής
Μάθησης: Σύγκριση και Ερμηνεία Αλγόριθμων
Ανίχνευσης Ανωμαλιών σε Δεδομένα Δικτύου»**

**Διπλωματική Εργασία η οποία υποβλήθηκε προς απόκτηση
Μεταπτυχιακού τίτλου σπουδών στα Πληροφοριακά Συστήματα
στο Πανεπιστήμιο Νεάπολις Πάφος.**

ΑΘΑΝΑΣΙΟΣ ΣΚΑΜΑΓΚΑΣ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ (16 Font Size, bold)

06/2026

Πνευματικά δικαιώματα

Copyright © Όνομα επίθετο φοιτητή, έτος κατάθεσης Διπλωματικής Εργασίας

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Η έγκριση της Διπλωματικής Εργασίας από το Πανεπιστήμιο Νεάπολις δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Πανεπιστημίου.

Πίνακας περιεχομένων

Abstract.....	8
1. Εισαγωγή.....	9
1.1 Το πρόβλημα της ανίχνευσης ανωμαλιών στα σύγχρονα δίκτυα.....	10
1.2 Προκλήσεις σχετικά με τα datasets και την αξιολόγηση μοντέλων IDS	10
1.3 Ο ρόλος της Μηχανικής Μάθησης και των Υβριδικών Τεχνικών	11
1.4 Η ανάγκη για Explainable AI στα συστήματα IDS	11
1.5 Το κενό στη βιβλιογραφία και η συμβολή της παρούσας εργασίας.....	12
1.6 Στόχοι της εργασίας.....	12
2. Βιβλιογραφική Ανασκόπηση.....	14
2.1. Εισαγωγή στα Συστήματα Ανίχνευσης Εισβολών και στην Ανίχνευση Ανωμαλιών	14
2.2. Machine Learning για Εντοπισμό Ανωμαλιών	15
2.2.1 Προβλήματα Γενίκευσης.....	16
2.2.2 Προβλήματα Feature Space & Preprocessing	16
2.2.3 Flow Based vs Payload-based ML	17
2.2.4 Χρήση Deep Learning για IoT IDS	17
2.3 Ζητήματα Ποιότητας Δεδομένων και Standardization στα Network Intrusion Datasets.....	17
2.4 Explainable Artificial Intelligence (XAI) στο Χώρο της Ανίχνευσης Ανωμαλιών	19
2.5 Cross-Dataset Generalization και Benchmarking	20
2.6 Ερευνητικά Κενά και Προοπτικές.....	22
2.6.1 Περιορισμοί των Υφιστάμενων Datasets	22
2.6.2 Περιορισμοί Μεθοδολογίας και Απόδοσης ML Μοντέλων.....	23
2.6.3 Ελλιπής Ενσωμάτωση Explainable AI.....	24
2.6.4 Προβλήματα Αναπαραγωγιμότητας και Ελλιπές Processing.....	25
2.6.5 Έλλειψη Robustness και Adversarial Testing	25
2.7 Σύνοψη	26
3. Μεθοδολογία	26
3.1 Ερευνητικός Σχεδιασμός	27
3.2 Σύνολα Δεδομένων	29
3.2.1 CICIDS2017	29
3.2.2 UNSW-NB15	29
3.3 Επιλογή και Αντιστοίχιση Χαρακτηριστικών	30
3.4 Μηχανική Χαρακτηριστικών	31

3.5 Προεπεξεργασία Δεδομένων	32
3.6 Αντιμετώπιση Ανισορροπίας Κλάσεων	33
3.7 Μοντέλα Μηχανικής Μάθησης	33
3.7.1 Logistic Regression	34
3.7.2 Random Forest.....	34
3.7.3 LinearSVC	34
3.8 Σχεδιασμός Πειραμάτων	35
3.9 Μετρικές Αξιολόγησης.....	35
3.10 Ανάλυση Μετατόπισης Κατανομής Χαρακτηριστικών	36
3.11 Explainable Failure Analysis με SHAP.....	37
3.12 Υλοποίηση και Αναπαραγωγιμότητα	37
3.13 Περιορισμοί της Μεθοδολογίας	38
3.14 Σύνοψη Κεφαλαίου	39
4. Πειραματικά Αποτελέσματα & Ανάλυση	39
4.1 Συνοπτική Παρουσίαση Αποτελεσμάτων	40
4.2 Same-Dataset Αποτελέσματα	41
4.2.1 Πείραμα CIC→CIC.....	41
4.2.2 Πείραμα UNSW→UNSW.....	43
4.2.3 Συμπέρασμα Same-Dataset Πειραμάτων	44
4.3 Cross-Dataset Αποτελέσματα.....	45
4.3.1 Πείραμα CIC→UNSW.....	45
4.3.2 Πείραμα UNSW→CIC.....	47
4.3.3 Συμπέρασμα Cross-Dataset Πειραμάτων	48
4.4 Σύγκριση Same-Dataset και Cross-Dataset Απόδοσης	48
4.5 Ανάλυση Αποτυχίας ανά Μοντέλο.....	51
4.5.1 Logistic Regression	51
4.5.2 Random Forest.....	51
4.5.3 LinearSVC	52
4.6 Πρακτική Ερμηνεία των Αποτελεσμάτων.....	52
4.7 Σύνδεση με το Πρόβλημα του Dataset Shift	53
4.8 Σύνοψη Κεφαλαίου	53
5.Ανάλυση Αιτιών και Ερμηνεία Αποτελεσμάτων	54
5.2 Μετατόπιση Κατανομής Χαρακτηριστικών.....	55
5.3 Ο Ρόλος του RobustScaler.....	57
5.4 Ερμηνεία της Αποτυχίας CIC→UNSW	58

5.5 Ερμηνεία της Αποτυχίας UNSW→CIC	59
5.6 SHAP-Based Explainable Failure Analysis	59
5.7 SHAP Ανάλυση στα Same- Dataset Πειράματα	60
5.7.1 CIC→CIC	60
5.7.2 UNSW→UNSW	61
5.8 SHAP Ανάλυση στα Cross-Dataset Πειράματα	62
5.8.1 CIC→UNSW	62
5.8.2 UNSW→CIC	64
5.9 Σύγκριση SHAP Σημαντικότητας Μεταξύ Πειραμάτων	65
5.10 Σύνδεση Feature Shift και SHAP	66
5.11 Επιχειρησιακές Επιπτώσεις	67
5.12 Περιορισμοί της Ανάλυσης	68
5.13 Σύνοψη Κεφαλαίου	68
6. Συμπεράσματα και Μελλοντική Έρευνα	69
6.1 Βασικά Συμπεράσματα της Εργασίας	70
6.2 Ερμηνεία της Αποτυχίας Γενίκευσης	70
6.3 Επιστημονική και Πρακτική Συμβολή	71
6.4 Περιορισμοί της Εργασίας	72
6.5 Προτάσεις για Μελλοντική Έρευνα	73
6.6 Τελικό Συμπέρασμα	74
7. Βιβλιογραφία	75

Σελίδα Εγκυρότητας

Ονοματεπώνυμο Φοιτητή: Αθανάσιος Σκαμάγκας

Τίτλος Διπλωματικής Εργασίας: «Ανίχνευση Κυβερνοεπιθέσεων με Τεχνικές Μηχανικής Μάθησης: Σύγκριση και Ερμηνεία Αλγόριθμων Ανίχνευσης Ανωμαλιών σε Δεδομένα Δικτύου»

Η παρούσα Διπλωματική Εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση εξ αποστάσεως μεταπτυχιακού τίτλου στο Πανεπιστήμιο Νεάπολις και εγκρίθηκε στις 09/10/2025 από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή:

Πρώτος επιβλέπων (Πανεπιστήμιο Νεάπολις Πάφος).....[ονοματεπώνυμο, βαθμίδα, υπογραφή]

Μέλος Εξεταστικής Επιτροπής:[ονοματεπώνυμο, βαθμίδα, υπογραφή]

Μέλος Εξεταστικής Επιτροπής:[ονοματεπώνυμο, βαθμίδα, υπογραφή]

Η ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

Ο Αθανάσιος Σκαμάγκας, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Ανίχνευση Κυβερνοεπιθέσεων με Τεχνικές Μηχανικής Μάθησης: Σύγκριση και Ερμηνεία Αλγόριθμων Ανίχνευσης Ανωμαλιών σε Δεδομένα Δικτύου», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Ο Δηλών



Abstract

Η ταχεία αύξηση της κίνησης δεδομένων και ευρεία χρήση ετερογενών τεχνολογιών δικτύου καθιστούν την ανίχνευση κυβερνοεπιθέσεων μια βασική πρόκληση για σύγχρονα συστήματα κυβερνοασφάλειας. Αν και οι μέθοδοι Μηχανικής Μάθησης (ML) έχουν επιτύχει εντυπωσιακά αποτελέσματα στην ταξινόμηση κακόβουλων δραστηριοτήτων, τα περισσότερα μοντέλα παρουσιάζουν σοβαρή αδυναμία σε αυτό που λέμε «γενίκευση», δηλαδή στη διαδικασία κατά την οποία τα μοντέλα αυτά εφαρμόζονται σε δεδομένα διαφορετικά από αυτά με τα οποία εκπαιδεύτηκαν. Το φαινόμενο αυτό, γνωστό και ως Cross-Dataset Generalization Gap, έχει καταγραφεί εκτενώς στη διεθνή βιβλιογραφία, με μελέτες να αναφέρουν χαρακτηριστικές πτώσεις απόδοσης άνω του 50-90% όταν μοντέλα εκπαιδευμένα σε CICIDS2017 ή UNSW-NB15 εφαρμόζονται σε IoT datasets όπως IoT-23, ToN-IoT ή BoT-IoT. Επιπλέον, η έλλειψη μίας ενιαίας διαδικασίας προεπεξεργασίας, η ανομοιογένεια των χαρακτηριστικών (features) και τα προβλήματα Labeling δυσκολεύουν περαιτέρω την αξιολόγηση και ακόμη και τη σύγκριση ML τεχνικών.

Η παρούσα εργασία εξετάζει εις βάθος το πρόβλημα της ανίχνευσης ανωμαλιών σε πολλαπλά και ανομοιογενή σύνολα δεδομένων, εστιάζοντας στη συγκριτική αξιολόγηση αλγορίθμων ML και στην ερμηνεία των μοντέλων (XAI). Αναλύοντας σύγχρονες τεχνικές ταξινόμησης και ανίχνευσης ανωμαλιών σε flow-based traffic, με χρήση δημοφιλών datasets όπως CICIDS2017, UNSW-NB15, BoT-IoT και IoT-23. Παράλληλα διερευνάται ο ρόλος του Explainable AI ως εργαλείο κατανόησης της συμπεριφοράς των μοντέλων συγκεκριμένα σε περιπτώσεις όπου αποτυγχάνουν να «γενικεύσουν» σε νέα περιβάλλοντα.

Στόχος της εργασίας είναι η ανάπτυξη και η αξιολόγηση ενός πλαισίου Explainable Cross-Dataset Intrusion Detection, το οποίο αξιοποιεί τεχνικές ML και XAI με στόχο:

- a) Τη μείωση του cross-dataset performance gap
- b) Την κατανόηση των παραγόντων που επηρεάζουν τη γενίκευση των μοντέλων και
- c) Την ενίσχυση της αξιοπιστίας τους μέσω ερμηνεύσιμων εξηγήσεων.

Τα αποτελέσματα φιλοδοξούμε να συμβάλουν στην κατανόηση των περιορισμών των ML-based IDS και τη διαμόρφωση πιο ανθεκτικών, επεκτάσιμων και επεξηγήσιμων μηχανισμών ανίχνευσης κυβερνοεπιθέσεων.

1. Εισαγωγή

Η συνεχής ψηφιοποίηση κοινωνικών, οικονομικών και βιομηχανικών υποδομών έχει οδηγήσει σε μία εκρηκτική αύξηση της κίνησης δεδομένων και της εξάρτησης από δικτυακές υπηρεσίες. Η ίδια εξέλιξη συνοδεύτηκε από αντίστοιχη αύξηση σε κλίμακα, συχνότητα και πολυπλοκότητα κυβερνοεπιθέσεων, οι οποίες στοχεύουν τόσο παραδοσιακά δίκτυα όσο και σύγχρονες πλατφόρμες Internet of Things (IoT), βιομηχανικούς κόμβους Industry 4.0/5.0, και κατανεμημένα υπολογιστικά περιβάλλοντα (*Emanuel Kryszton et al., 2024; Makridis, 2024*). Η έντονη αυτή μεταβολή έχει καταστήσει επιτακτική τη μετάβαση από κλασσικά, signature-based Intrusion Detection Systems (IDS) σε data-driven λύσεις που βασίζονται σε Machine Learning (ML), Deep Learning (DL) και Explainable Artificial Intelligence (XAI).

Τα σύγχρονα ML-based IDS καταφέρνουν να εντοπίζουν επιθέσεις που δεν έχουν καταγραφεί σε βάσεις υπογραφών, να ανιχνεύσουν zero-day threats, αλλά και να αναλύουν τεράστιους όγκους δεδομένων σε πραγματικό χρόνο (*Rodriguez et al., 2022, Vibhute, 2024*). Ωστόσο, η αποτελεσματικότητα των μοντέλων αυτών επηρεάζεται σημαντικά από ζητήματα όπως η ποιότητα και η αντιπροσωπευτικότητα των datasets (*Rosay et al., 2022, Alhowaide & Alsmadi, 2022*), η ικανότητα γενίκευσης σε διαφορετικά περιβάλλοντα (cross-dataset generalization) (*Prazeres, 2023*), η ευπάθεια σε adversarial επιθέσεις (*Catillo et al., 2023*) και η έλλειψη διαφάνειας στον τρόπο λήψης αποφάσεων των black-box μοντέλων (*Osvaldo Arreche et al., 2023*). Αυτές οι προκλήσεις δημιουργούν ένα νέο ερευνητικό πεδίο που συνδυάζει τεχνικές ανίχνευσης ανωμαλιών με μεθοδολογίες Explainable AI, προκειμένου να επιτευχθεί υψηλή ακρίβεια

χωρίς να θυσιάζεται η ερμηνευσιμότητα, η αξιοπιστία και η ανθρώπινη εποπτεία (Rudresh et al., 2023, Galzarossa et al., 2025).

1.1 Το πρόβλημα της ανίχνευσης ανωμαλιών στα σύγχρονα δίκτυα

Οι σύγχρονες κυβερνοεπιθέσεις χαρακτηρίζονται από μεγάλη πολυπλοκότητα, αυτοματοποίηση και δυνατότητα παραλλαγής, στοιχεία που δυσχεραίνουν σημαντικά τη χρήση κλασικών detection κανόνων. Επιθέσεις όπως DDoS, botnet, ransomware, brute-force και IoT device hijacking απαιτούν τεχνικές δυναμικής ανάλυσης συμπεριφοράς και deep feature extraction (Sarhan et al., 2024). Ιδιαίτερα στο IoT, συσκευές με περιορισμένη υπολογιστική ισχύ παράγουν ανομοιογενή δεδομένα και εκτίθενται συχνά χωρίς επαρκή προστασία, γεγονός που τις καθιστά ιδανικό στόχο για επιθέσεις μεγάλης κλίμακας (Krzyszton et al., 2024, Prazeres, 2023).

Παράλληλα, επιθέσεις που αναπτύσσονται σε κινητικά περιβάλλοντα, σε βιομηχανικά δίκτυα Industry 5.0 ή ακόμα και στο metaverse (Deepika Singh et al., 2024) απαιτούν προηγμένες τεχνικές adaptive learning ώστε να ανιχνεύονται αποκλίσεις από τη φυσιολογική συμπεριφορά του συστήματος.

Σε αυτό το πλαίσιο, η ανίχνευση ανωμαλιών γίνεται κρίσιμη προσέγγιση για την αποτροπή περιστατικών που δεν συμπεριλαμβάνονται σε signature databases (More et al., 2024, Ajagbe et al., 2024).

1.2 Προκλήσεις σχετικά με τα datasets και την αξιολόγηση μοντέλων IDS

Ένα κεντρικό ζήτημα που αναγνωρίζεται επανειλημμένα στη βιβλιογραφία είναι το πρόβλημα της αντιπροσωπευτικότητας των datasets. Πολλά από τα δημοφιλή σύνολα δεδομένων (π.χ. NSL-KDD, CICIDS2017, UNSW-NB15) παρουσιάζουν σημαντικούς περιορισμούς: είναι παλιά, ανισόρροπα, μη αντιπροσωπευτικά του σημερινού traffic ή κατασκευασμένα υπό τεχνητές συνθήκες (Rosay et al., 2022, Wali Syed et al., 2024). Επιπλέον, η χρήση ενός dataset για training και διαφορετικού για testing (cross-dataset evaluation) οδηγεί συχνά σε μεγάλη πτώση της απόδοσης, δείχνοντας ότι τα περισσότερα μοντέλα δεν γενικεύουν σωστά (Prazeres, 2023, Krzyszton et al., 2024).

Η ανάγκη για ρεαλιστικά και flow-based datasets γίνεται όλο και πιο επιτακτική. Νέες προσπάθειες, όπως το CICIoT2023 το οποίο είναι πλήρως προσανατολισμένο σε real-time IoT traffic (Pinto Neto et al., 2023), επιχειρούν να αντιμετωπίσουν τα προβλήματα

αυτά, ενώ παράλληλα αναδεικνύουν τη σημασία της αναπαράστασης των επιθέσεων σε επίπεδο flows αντί packets (Rodriguez et al., 2022).

Ένα δεύτερο πρόβλημα είναι η ανισορροπία των δεδομένων, η οποία οδηγεί τα μοντέλα σε bias προς την πλειοψηφική κλάση. Έρευνες όπως αυτή των Alhowaide & Alsmadi (2022) τονίζουν ότι τα balanced datasets είναι προϋπόθεση για αξιόπιστα αποτελέσματα.

1.3 Ο ρόλος της Μηχανικής Μάθησης και των Υβριδικών Τεχνικών

Οι σύγχρονες προσεγγίσεις ML/DL περιλαμβάνουν CNNs, LSTMs, Autoencoders, Random Forest, XGBoost, Transformers και υβριδικές αρχιτεκτονικές (Donkol & Gad, 2018 – More et al., 2024). Μεγάλο ενδιαφέρον παρουσιάζει η χρήση θβριδικών λύσεων, όπως:

- LSTM-RNN + swarm intelligence optimization που επιτυγχάνει υψηλή ακρίβεια σε temporal traffic patterns (Donkol & Gad, 2018)
- CNNs για φιλτράρισμα συμπεριφορών μοτίβων σε πολυδιάστατα δικτυακά χαρακτηριστικά (Vibhule et al., 2024)
- AutoML frameworks που αυτοματοποιούν τον σχεδιασμό δικτυακών pipelines (Yang, 2023)
- Federated learning για προστασία δεδομένων και edge συνεργασία (Krzyszton et al., 2024).

Ωστόσο, η μεγάλη υπολογιστική πολυπλοκότητα και ο κίνδυνος overfitting, ειδικά σε με ρεαλιστικά datasets, καθιστούν αναγκαία την αξιολόγηση σε cross-dataset συνθήκες.

1.4 Η ανάγκη για Explainable AI στα συστήματα IDS

Παρότι deep learning μοντέλα πετυχαίνουν εντυπωσιακά ποσοστά ακρίβειας, η έλλειψη διαφάνειας περιορίζει τη χρησιμότητα τους περιβάλλοντα όπου απαιτείται ανθρώπινη λήψη αποφάσεων, συμμόρφωση, auditing και real-time trust (Rudresh et al., 2023).

Πρόσφατες μελέτες τονίζουν ότι το XAI:

- μπορεί να αποκαλύψει bias σε datasets (Wali Syed et al., 2025)
- βοηθά μη ειδικούς να κατανοήσουν τις αποφάσεις ενός IDS (Osvaldo Arreche et al., 2023)
- ενισχύει την ανθεκτικότητα απέναντι σε adversarial επιθέσεις (Castillo et al., 2023)

- επιτρέπει επεξηγήσεις σε βιομηχανικά περιβάλλοντα *Industry 5.0* όπου απαιτείται accountability (Makridis, 2024)
- λειτουργεί ως μηχανισμός αξιολόγησης ML μοντέλων με βάση τη διαφάνεια και όχι μόνο της ακρίβεια (Galzarossa et al., 2025)

Το XAI πλέον έχει καθιερωθεί ως βασικό συστατικό κάθε συστήματος κυβερνοασφάλειας που βασίζεται σε ML.

1.5 Το κενό στη βιβλιογραφία και η συμβολή της παρούσας εργασίας

Αν και υπάρχει έντονο ερευνητικό ενδιαφέρον για ML-based NIDS και για XAI, η πλειονότητα των εργασιών αξιολογεί τα μοντέλα σε ένα μόνο dataset, γεγονός που παράγει αποτελέσματα χωρίς πραγματικό νόημα για επιχειρησιακά περιβάλλοντα. Όπως υποδεικνύουν πρόσφατες εργασίες (Prazeres, 2023 – Krzyszton et al., 2024), η πτώση των επιδόσεων όταν ένα μοντέλο δοκιμάζεται σε διαφορετικά dataset είναι δραματική – συχνά άνω του 50%.

Επιπλέον, η χρήση XAI σε flow-based μοντέλα βρίσκεται ακόμη σε πρώιμο στάδιο, με τις περισσότερες μελέτες να εφαρμόζουν XAI γενικά, χωρίς να εξετάζουν το ποια χαρακτηριστικά είναι πραγματικά σημαντικά σε cross-dataset περιβάλλοντα, το πως εξηγούνται τα patterns ανωμαλιών στα flows, το ποια XAI frameworks (π.χ. SHAP, LIME, Integrated Gradients) λειτουργούν καλύτερα με δικτυακά χαρακτηριστικά και το αν οι XAI εξηγήσεις γενικεύουν μεταξύ datasets.

Έτσι, υπάρχει μια εμφανής ανάγκη για Explainable Cross-Dataset Machine Learning IDS, που να εξετάζει όχι μόνο την ακρίβεια, αλλά και τον τρόπο με τον οποίο οι εξηγήσιμες μέθοδοι αποκαλύπτουν patterns διαχρονικά και μεταξύ διαφορετικών κατανομών traffic.

1.6 Στόχοι της εργασίας

Η παρούσα διπλωματική εργασία έχει ως βασικό στόχο τη διερεύνηση της ικανότητας γενίκευσης μοντέλων Μηχανικής Μάθησης για την ανίχνευση κυβερνοεπιθέσεων σε διαφορετικά σύνολα δεδομένων δικτυακής κίνησης. Ειδικότερα, η εργασία εστιάζει στο πρόβλημα του **cross-dataset generalization**, δηλαδή στην αξιολόγηση του κατά πόσο ένα μοντέλο που εκπαιδεύεται σε ένα dataset μπορεί να διατηρήσει ικανοποιητική απόδοση όταν εφαρμόζεται σε ένα διαφορετικό dataset, με διαφορετική κατανομή

χαρακτηριστικών, διαφορετική μεθοδολογία συλλογής και διαφορετική αναπαράσταση δικτυακών ροών.

Σε αντίθεση με πολλές μελέτες που αξιολογούν μοντέλα μόνο σε train/test διαχωρισμούς του ίδιου dataset, η παρούσα εργασία εξετάζει τόσο same-dataset όσο και cross-dataset πειραματικά σενάρια. Με αυτόν τον τρόπο επιδιώκεται να διαπιστωθεί αν οι υψηλές επιδόσεις των ML-based IDS αντανακλούν πραγματική ικανότητα αναγνώρισης κακόβουλης συμπεριφοράς ή αν οφείλονται κυρίως στην προσαρμογή των μοντέλων στις ιδιαιτερότητες του dataset εκπαίδευσης.

Οι επιμέρους στόχοι της εργασίας είναι οι εξής:

1. **Να αξιολογηθεί η απόδοση παραδοσιακών supervised Machine Learning μοντέλων** σε προβλήματα δυαδικής ταξινόμησης δικτυακής κίνησης, όπου οι εγγραφές κατηγοριοποιούνται ως benign ή attack.
2. **Να συγκριθεί η απόδοση των μοντέλων σε same-dataset και cross-dataset συνθήκες**, χρησιμοποιώντας τα datasets CICIDS2017 και UNSW-NB15.
3. **Να εξεταστεί η επίδραση της μετατόπισης κατανομής χαρακτηριστικών** μεταξύ διαφορετικών datasets στην απόδοση των μοντέλων ανίχνευσης εισβολών.
4. **Να χρησιμοποιηθεί κοινό feature space** μεταξύ CICIDS2017 και UNSW-NB15, ώστε να είναι δυνατή η μεθοδολογικά ελεγχόμενη cross-dataset αξιολόγηση.
5. **Να αναλυθεί η απόδοση των μοντέλων με μετρικές κατάλληλες για ανισόρροπα datasets**, όπως precision, recall, F1-score, balanced accuracy και confusion matrices.
6. **Να εφαρμοστεί SHAP-based explainable failure analysis** στο Random Forest, με σκοπό την ερμηνεία των αποφάσεων του μοντέλου και την κατανόηση των λόγων αποτυχίας του σε cross-dataset σενάρια.
7. **Να αναδειχθεί ο ρόλος της προεπεξεργασίας, της αντιστοίχισης χαρακτηριστικών και της εναρμόνισης του feature space** ως κρίσιμων παραγόντων για την αξιοπιστία των ML-based IDS.

Με βάση τους παραπάνω στόχους, η εργασία δεν επιχειρεί να κατασκευάσει ένα πλήρες real-time IDS ούτε να συγκρίνει σύνθετες αρχιτεκτονικές Deep Learning. Αντίθετα, επικεντρώνεται σε μια ελεγχόμενη και αναπαραγώγιμη πειραματική διαδικασία, η οποία εξετάζει αν κλασικά μοντέλα Μηχανικής Μάθησης μπορούν να γενικεύσουν πέρα από το dataset στο οποίο εκπαιδεύτηκαν. Η προσέγγιση αυτή επιτρέπει την ανάδειξη ενός κρίσιμου περιορισμού της υπάρχουσας βιβλιογραφίας: η υψηλή απόδοση σε ένα

μεμονωμένο dataset δεν συνεπάγεται απαραίτητα επιχειρησιακή αξιοπιστία σε διαφορετικά δικτυακά περιβάλλοντα.

2. Βιβλιογραφική Ανασκόπηση

2.1. Εισαγωγή στα Συστήματα Ανίχνευσης Εισβολών και στην Ανίχνευση Ανωμαλιών

Η ασφάλεια δικτύων αποτελεί κρίσιμο πυλώνα για σύγχρονα πληροφοριακά συστήματα, καθώς οι κυβερνοεπιθέσεις έχουν εξελιχθεί τόσο σε όγκο όσο και σε πολυπλοκότητα, καθιστώντας αναγκαίες νέες μεθόδους εντοπισμού και αντιμετώπισης απειλών. Τα Networks Intrusion Detection Systems (NIDS) συνιστούν βασικά εργαλεία σε αυτή την προσπάθεια, καθώς επιτρέπουν τη συνεχή παρακολούθηση της δικτυακής κίνησης και την αναγνώριση ανωμαλιών ή κακόβουλων ενεργειών (Moustafa et al. 2019). Σε αντίθεση με τα παραδοσιακά συστήματα ανίχνευσης που βασίζονται σε κανόνες ή υπογραφές οι σύγχρονες προσεγγίσεις στηρίζονται σε τεχνικές Μηχανικής Μάθησης (ML) και Τεχνητής Νοημοσύνης (AI), οι οποίες εμφανίζουν αυξημένη ικανότητα ανίχνευσης νέων ή εξελιγμένων επιθέσεων μέσα από την ανάλυση μοτίβων της κυκλοφορίας.

Η σχετική βιβλιογραφία διακρίνει δύο κυρίαρχες κατευθύνσεις ως προς την αναπαράσταση των δεδομένων: packet-based και flow-based ανίχνευση. Η flow-based προσέγγιση έχει αποκτήσει ιδιαίτερη σημασία, ειδικά σε περιβάλλοντα IoT και μεγάλης κλίμακας συστήματα, λόγω της δυνατότητας της για αποτελεσματική, χαμηλού κόστους και ταχύτατη επεξεργασία (Santos et al., 2021, Velan, 2018). Τα flows παρέχουν συνοπτική εικόνα επικοινωνιών μεταξύ endpoints και επιτρέπουν την εξαγωγή στατιστικών χαρακτηριστικών που σχετίζονται με πρωτόκολλα, μετρητές πακέτων, χρόνους διασύνδεσης κ.λ.π.. Ωστόσο, η απουσία payload ανάλυσης αποτελεί σημαντικό περιορισμό για επιθέσεις που εκδηλώνονται στο περιεχόμενο των πακέτων, όπως SQL injection ή web-based exploits (Rosay et al., 2022).

Η έλλειψη τυποποίησης στα διαθέσιμα datasets αποτελεί επίσης καθοριστικό εμπόδιο στη γενίκευση των ML μοντέλων. Ιστορικά, datasets όπως το KDD99 και NSL-KDD αναγνωρίστηκαν ως ανεπαρκή λόγω αλληλοεπικαλυπτόμενων εγγράφων, μη ρεαλιστικών προτύπων κίνησης και τεχνητών επιθέσεων (Pekar et al., 2024). Νεότερα σύνολα, όπως τα CICIDS2017, CICDDoS2019, UNSW-NB15 και IoT-23 πρόσφεραν

σημαντική πρόοδο, ενσωματώνοντας πιο ρεαλιστικά χαρακτηριστικά ροής και ποικιλία επιθέσεων. Ωστόσο, αρκετές μελέτες παρουσίασαν προβλήματα αξιοπιστίας και συνέπειας, όπως λάθος χρονικά σήματα, μη ευθυγραμμισμένα labels, ατελή flows ή ελλιπή χαρακτηριστικά (Rodriguez et al., 2022- Castillo et al., 2023).

Σημαντικό ζήτημα αποτελεί η αδυναμία cross-dataset generalization. Πλήθος μελετών απέδειξε ότι μοντέλα που εκπαιδεύονται σε ένα dataset παρουσιάσουν κατακόρυφη πτώση στην απόδοση όταν εφαρμόζονται σε ένα dataset παρουσιάζουν κατακόρυφη πτώση στην απόδοση όταν εφαρμόζονται σε άλλο, ακόμη κι όταν οι κατηγορίες επιθέσεων είναι παρόμοιες (Ajagbe et al., 2024- Makridis & colleagues, 2023). Αυτό οφείλεται σε διαφορετικές μεθοδολογίες συλλογής δεδομένων, ετερογένεια χαρακτηριστικών, ασυνέπειες στις ροές, καθώς και επιθέσεις που εφαρμόζονται με διαφορετικούς ρυθμούς και μοτίβα σε κάθε dataset (Rosay et al., 2022). Ένα από τα πιο κρίσιμα κενά που αναδεικνύει η σύγχρονη βιβλιογραφία είναι η έλλειψη τυποποιημένου, ενιαίου χώρου χαρακτηριστικών (feature space) που να επιτρέπει αξιόπιστη διασταυρούμενη αξιολόγηση και εκπαίδευση μοντέλων (LycosTand dataset, 2024).

Συμπερασματικά, η σχετική βιβλιογραφία αναδεικνύει:

- Την ανάγκη για unified feature extraction
- Τη σημασία των flow-based χαρακτηριστικών λόγω αποδοτικότητας
- Την κρίσιμη αδυναμία του ML να γενικεύει σε πολλαπλά datasets
- Την έλλειψη ερμηνευσιμότητας (Explainable AI) στα αποτελέσματα
- Την ανάγκη για cross-dataset αξιολόγηση με XAI, που αποτελεί και το κεντρικό σημείο καινοτομίας της παρούσας εργασίας.

2.2. Machine Learning για Εντοπισμό Ανωμαλιών

Η χρήση τεχνικών Μηχανικής Μάθησης (ML) στον εντοπισμό κυβερνοεπιθέσεων έχει εξελιχθεί σε κυρίαρχη κατεύθυνση, καθώς τα σύγχρονα δίκτυα παράγουν τεράστιους όγκους δεδομένων και εμφανίζουν πολύπλοκα μοτίβα συμπεριφοράς που ξεπερνούν τις δυνατότητες των συστημάτων βασισμένων σε κανόνες (Berman et al., 2019). Η Μηχανική Μάθηση επιτρέπει την αυτόματη αναγνώριση προτύπων σε flows δικτυακής κίνησης, εντοπίζοντας αποκλίσεις από τη φυσιολογική συμπεριφορά (Chaabouni et al., 2019).

Οι αλγόριθμοι Μηχανικής Μάθησης που εμφανίζουν καλύτερη απόδοση στα NIDS περιλαμβάνουν δέντρα απόφασης, Random Forest, Support Vector Machines, k-NN, καθώς και αλγόριθμους «βαθιάς μάθησης» (DL) όπως DNNs, CNNs και LSTMs (Ajagbe et al., 2024- Vibhute et al., 2024- Makridis et al., 2023). Κάθε αλγόριθμος παρουσιάζει διαφορετικά πλεονεκτήματα:

- RF και ensembles επιτυγχάνουν υψηλή ακρίβεια και F1-score σε datasets όπως UNSW-NB15 και CICIDS2017 (Ajagbe et al., 2024- More et al., 2024).
- CNN και LSTM παρουσιάζουν εξαιρετική απόδοση στην ανίχνευση botnet συμπεριφορών σε BoT-IoT και IoT-23 (Vibhute et al., 2024- Austin, 2021).
- Hybrid μοντέλα (π.χ. CNN & RF) αξιοποιούν τα πλεονεκτήματα και των δύο κόσμων, επιτυγχάνοντας $F1 > 0.95$ σε πολλές κατηγορίες επιθέσεων (BoT-IoT studies, 2022-2024).

Παρ' όλ' αυτά, η βιβλιογραφία αναδεικνύει σημαντικά προβλήματα:

2.2.1 Προβλήματα Γενίκευσης

Η απόδοση των μοντέλων παρουσιάζει μεγάλη πτώση όταν εφαρμοστούν σε διαφορετικό dataset.

Για παράδειγμα:

- Ένα μοντέλο που εκπαιδεύτηκε σε IoT-23 αποτυγχάνει να ανιχνεύσει ανωμαλίες στο MQTT-IoT-IDS2020 (Makridis et al., 2023).
- Μοντέλα βαθιάς μάθησης (DL) εμφανίζουν cross-dataset F1-score κάτω από 0.30 σε payload-based επιθέσεις (UM-NIDS, 2024).
- Ακόμη και RF classifiers αποτυγχάνουν όταν οι ροές προέρχονται από διαφορετικά περιβάλλοντα λήψης (Pekar et al., 2024).

Το πρόβλημα της γενίκευσης αποτελεί ένα από τα σημαντικότερα ζητήματα στην επιστήμη των IDS.

2.2.2 Προβλήματα Feature Space & Preprocessing

Αναλύσεις σε CICIDS2017 αποκάλυψαν:

- Λάθη στη ροή χαρακτηριστικών (Rodriguez et al., 2022)
- Ασυνεπείς χρονικές σφραγίδες
- Ασύμβατα Labels
- Κακό Alignment μεταξύ raw PCAP και extracted flows.

Η εργασία *LycosTand* (2024) προτείνει επανεξαγωγή χαρακτηριστικών, δημιουργώντας το LYCOS-IDS2017 με σημαντικές βελτιώσεις στην απόδοση όλων των ML μοντέλων. Αυτό επιβεβαιώνει ότι η απόδοση των μοντέλων εξαρτάται περισσότερο από την ποιότητα του dataset παρά από τον ίδιο τον αλγόριθμο ML.

2.2.3 Flow Based vs Payload-based ML

Η βιβλιογραφία δείχνει ότι:

- Flow-based ML είναι γρήγορο και ελαφρύ (*Santos et al., 2021*).
- Payload-based ML ανιχνεύει πιο εξελιγμένες επιθέσεις που δεν εμφανίζουν flow anomalies (SQLi, XSS) (*UM-NIDS, 2024*).

Ένα εντυπωσιακό εύρημα δείχνει ότι:

- **Flow-based ML** πέτυχε $F1 = 0.97$ στο ίδιο dataset, αλλά μόλις 0.31 σε cross-dataset test.
- **Payload-based ML** πέτυχε $F1 = 0.86$ cross-dataset, δείχνοντας πολύ καλύτερη μεταφορά γνώσης.

2.2.4 Χρήση Deep Learning για IoT IDS

Στην IoT βιβλιογραφία, όπου τα flows είναι μικρά και οι συσκευές περιορισμένων πόρων, χρησιμοποιούνται:

- GRU (*Mothukuri et al., 2022*)
- LSTM σε MQTT flows (*Hindy et al., 2020*)
- CNN σε IoT-23 (*Austin, 2021*)
- Federated Learning για κατανεμημένα IDS (*Makridis et al., 2023*)

Ωστόσο, όλες οι μελέτες αναφέρουν χαμηλή διασταυρούμενη γενίκευση.

2.3 Ζητήματα Ποιότητας Δεδομένων και Standardization στα Network Intrusion Datasets

Η ποιότητα των διαθέσιμων datasets για την ανίχνευση κυβερνοεπιθέσεων αποτελεί ίσως τη σημαντικότερη μεθοδολογική πρόκληση στον χώρο του Machine Learning-based Intrusion Detection. Από το 2022 και μετά, η διεθνής βιβλιογραφία συγκλίνει στο συμπέρασμα ότι η απόδοση των μοντέλων ML επηρεάζει δραματικά όχι μόνο από την επιλογή αλγορίθμου, αλλά κυρίως από την αξιοπιστία και συνέπεια των δεδομένων που χρησιμοποιούνται κατά την εκπαίδευση και αξιολόγηση (*Rosay et al., 2022-Rodriguez et al., 2022*). Η έρευνα δείχνει σταθερά ότι ακόμη και τα πλέον εξελιγμένα νευρωνικά

δίκτυα, όταν εκπαιδεύονται σε ατελή ή εσφαλμένα διαμορφωμένα σύνολα δεδομένων, οδηγούν σε υψηλή φαινομενική ακρίβεια, η οποία ωστόσο δεν αντικατοπτρίζει πραγματικές συνθήκες δικτυακής κίνησης.

Ένα από τα σοβαρότερα προβλήματα είναι η ασυμβατότητα μεταξύ raw packet captures (PCAP) και των flow-based datasets που παράγονται μέσω εργαλείων όπως το CICFlowMeter, το Zeek και άλλους flow exporters. Στο CICIDS2017, για παράδειγμα, έχει τεκμηριωθεί ότι η διαδικασία εξαγωγής ροών οδηγεί συχνά σε λανθασμένα timestamps, ασυνέχειες στη σειρά πακέτων και ελλιπή συσχέτιση μεταξύ διευθύνσεων, πρωτοκόλλων και flags (Rosay et al., 2022). Ως αποτέλεσμα, στατιστικά στοιχεία όπως η διάρκεια της ροής, το πλήθος πακέτων και οι ρυθμοί μετάδοσης προκύπτουν ανακριβή. Τα μοντέλα ML τελικά «μαθαίνουν» πρότυπα που δεν αντιστοιχούν σε πραγματικές επιθέσεις αλλά σε τεχνουργήματα του τρόπου καταγραφής και εξαγωγής. Η πρόσφατη εργασία *LycosTand* (2023-2024) κατέδειξε ότι η εκ νέου εξαγωγή χαρακτηριστικών από τα raw PCAP αρχεία, με αυστηρό έλεγχο συνέπειας, οδηγεί σε βελτιωμένη ακρίβεια και αξιοπιστία, αποδεικνύοντας ότι η αρχική μορφή πολλών δημοφιλών datasets είναι ελλιπής.

Σημαντικό ζήτημα αποτελεί επίσης η ακρίβεια του Labeling. Πολλά datasets δημιουργούνται με την υπόθεση ότι κάθε χρονική περίοδος που αντιστοιχεί στην ενεργοποίηση ενός script επίθεσης είναι εξολοκλήρου κακόβουλη, ενώ χρονικά τμήματα όπου δεν έχει ενεργοποιηθεί επίθεση θεωρούνται a priori benign. Αυτό οδηγεί σε ελλιπή ή λανθασμένη σήμανση, ιδιαίτερα σε περιπτώσεις όπου benign traffic συνυπάρχει με background activity την ίδια χρονική στιγμή. Το IoT-23 αποτελεί χαρακτηριστικό παράδειγμα: αρκετές ροές που χαρακτηρίζονται ως «malicious» αφορούν κανονική δικτυακή λειτουργία συσκευών. Άλλες, αντίστροφα, έχουν μείνει αταξινομήτες ενώ ανήκουν σε φάσεις command-and-control. Αυτή η μορφή labeling noise δημιουργεί φαινομενικά υψηλά ποσοστά ακρίβειας σε supervised αλγόριθμους, αλλά καταρρέει πλήρως όταν τα μοντέλα δοκιμάζονται σε ανεξάρτητα datasets.

Επιπλέον, η έλλειψη standardization στη διαδικασία εξαγωγής χαρακτηριστικών δημιουργεί ουσιαστικά ασύμβατα feature spaces μεταξύ datasets. Για παράδειγμα, το CICFlowMeter παράγει πάνω από 80 χαρακτηριστικά ανά ροή, με συγκεκριμένη ονομασία και μονάδες μέτρησης. Το Zeek, αντιθέτως, ακολουθεί μία εντελώς διαφορετική φιλοσοφία, εστιάζοντας σε TCP state transitions, επιτρέποντας representation με strings (όπως οι “SHF” flags). Άλλα datasets, όπως το UNSW-NB15 χρησιμοποιούν ιδιαίτερα τεχνικά χαρακτηριστικά, όπως entropy metrics, spkts και

dbytes, τα οποία δεν υφίστανται σε IoT datasets. Αυτή η απόλυτη ετερογένεια καθιστά αδύνατη την επανεκπαίδευση ή τη διασταυρωμένη απόδοση μοντέλων μεταξύ datasets χωρίς ένα ενιαίο preprocessing framework.

Έτσι, ο χώρος έχει ανάγκη από ενιαίες διαδικασίες καθαρισμού, ελέγχου συνέπειας, επιλογής χαρακτηριστικών και labeling. Προσπάθειες όπως το *UM-NIDS (2024)* επιδιώκουν να καλύψουν αυτό το κενό, συνθέτοντας multimodal datasets με ενοποιημένα χαρακτηριστικά, συνεπείς ορισμούς επιθέσεων και pipeline αυτόματης εναρμόνισης. Μέσα σε αυτό το πλαίσιο, η ποιότητα του dataset αποτελεί κρίσιμο παράγοντα επιτυχίας και ίσως το σημαντικότερο ανεξερεύνητο πεδίο στην έρευνα ML-based IDS.

2.4 Explainable Artificial Intelligence (XAI) στο Χώρο της Ανίχνευσης Ανωμαλιών

Η εφαρμογή Explainable Artificial Intelligence (XAI) στον τομέα της ανίχνευσης κυβερνοεπιθέσεων έχει εξελιχθεί σε έναν από τους σημαντικότερους και ταχύτερα αναπτυσσόμενους ερευνητικούς άξονες. Η βασική αφορμή αφορά την αδυναμία των συστημάτων IDS να εξηγήσουν τις αποφάσεις τους, ειδικά όταν στηρίζονται σε βαθιά νευρωνικά δίκτυα ή ensemble μοντέλα. Αυτά τα «μαύρα κουτιά» συχνά πετυχαίνουν εντυπωσιακή ακρίβεια σε benchmarks, αλλά δεν παρέχουν κανέναν μηχανισμό ερμηνείας των προβλέψεων τους, ούτε επιτρέπουν σε αναλυτές ασφαλείας να κατανοήσουν αν μια ταξινόμηση βασίζεται σε θεμελιώσεις ιδιότητες του traffic ή σε θόρυβο του dataset.

Η ανάγκη για XAI ενισχύεται από τους περιορισμούς των πραγματικών περιβαλλόντων. Στη βιομηχανία και στις κρίσιμες υποδομές, η άμεση αιτιολόγηση μιας ανίχνευσης αποτελεί προϋπόθεση τόσο για την επιχειρησιακή εμπιστοσύνη όσο και για τη συμμόρφωση με κανονιστικά πλαίσια (GDPR, NIS2). Τα συστήματα XAI επιτρέπουν στους αναλυτές να κατανοήσουν γιατί ένας ταξινομητής θεωρεί ότι μια ροή είναι κακόβουλη, ποια χαρακτηριστικά ήταν καθοριστικά και πως αυτά διαφοροποιούνται σε σχέση με το benign baseline.

Οι πλέον παραγωγικές τεχνικές XAI είναι οι SHAP, LIME και οι Counterfactual Explanations. Το SHAP, βασισμένο στη θεωρία παιγνίων, δίνει λεπτομερή ανάλυση της συμβολής κάθε χαρακτηριστικού στην τελική απόφαση του μοντέλου. Στα πρόσφατα paper *E-XAI frameworks (2024)*, έχει φανεί ότι οι SHAP values μπορούν να αποκαλύψουν patterns επιθέσεων που δεν είναι εύκολα ορατά με παραδοσιακά εργαλεία. Για παράδειγμα, σε επιθέσεις Mirai σε flow-based datasets, οι SHAP διαπιστώσεις

αποκαλύπτουν ότι η σημαντικότητα της διάρκειας ροής και του αριθμού μικρών SYN πακέτων είναι πολύ μεγαλύτερη απ' ό,τι εκτιμάται από απλές μετρικές όπως η μέση ταχύτητα μετάδοσης.

Αντίστοιχα, το LIME επιτρέπει τη δημιουργία τοπικών, ανθρώπινα κατανοητών μοντέλων γύρω από μία συγκεκριμένη πρόβλεψη. Στο πλαίσιο των web-based επιθέσεων ή των IoT κακόβουλων patterns, το LIME μπορεί να δείξει πως μικρές διαφοροποιήσεις στα χαρακτηριστικά τροποποιούν τη συμπεριφορά του ταξινομητή.

Ιδιαίτερα σημαντική εξέλιξη αποτελεί η εφαρμογή XAI στον χώρο των flow-based IDS, όπως αποδεικνύουν τα IoT-Flow IDS και Industry 5.0 IDS papers (2023). Σε αυτά τα συστήματα, η εξήγηση δεν αφορά μεμονωμένα πακέτα αλλά στατιστικές ροών. Το XAI, συνεπώς επιτρέπει ανάλυση υψηλότερου επιπέδου. Για παράδειγμα, ένας SHAP χάρτης μπορεί να καταδείξει ότι το χαρακτηριστικό “flow_packets” υπερέχει ως δείκτης anomaly στις slow DDoS επιθέσεις, ενώ το “packet_interarrival_variance” είναι πιο σημαντικό στις reconnaissance σάρωσης.

Τέλος, μια από τις πιο καινοτόμες χρήσεις του XAI στα IDS είναι ο εντοπισμός dataset shift. οι εξηγήσεις αποκαλύπτουν πως αλλάζει η σημασία των χαρακτηριστικών όταν το ίδιο μοντέλο δοκιμάζεται σε διαφορετικό dataset. Για παράδειγμα, ένα μοντέλο εκπαιδευμένο σε UNSW-NB15 μπορεί να βασίζεται στη «source_port_entropy» ως κρίσιμο χαρακτηριστικό, ενώ στο CICIDS2017 το ίδιο χαρακτηριστικό μέσω SHAP εμφανίζεται αμελητέο, αποδεικνύοντας ότι οι δύο πηγές δεδομένων έχουν βαθιά διαφορετικές στατιστικές ιδιότητες. Αυτή η δυνατότητα καθιστά το XAI χρήσιμο όχι μόνο για ερμηνεία αλλά και για βελτίωση της γενίκευσης, διόρθωση feature spaces και σχεδιασμό unified pipelines.

2.5 Cross-Dataset Generalization και Benchmarking

Η διασταυρωμένη γενίκευση (cross-dataset generalization) αποτελεί ίσως την πιο κρίσιμη και ταυτόχρονα πιο παραμελημένη παράμετρο στην έρευνα για ML-based IDS. η πλειονότητα των εργασιών αξιολογεί τα μοντέλα χρησιμοποιώντας ένα μόνο dataset ή, στη βέλτιστη περίπτωση, διαφορετικές τομές του ίδιου dataset. Ωστόσο, τα πραγματικά περιβάλλοντα δικτύωσης είναι πολύ πιο μεταβλητά, συχνά παρουσιάζοντας traffic distributions που διαφέρουν ριζικά από τα artificial benchmarks που δημιουργούνται σε εργαστηριακές συνθήκες.

Από το 2022 και μετά, σχεδόν όλες οι σημαντικές μελέτες (*Makridis et al., 2023-IoT-Flow IDS-E-XAI-UM-NIDS 2024*) καταλήγουν στο ίδιο συμπέρασμα, ότι δηλαδή τα μοντέλα ανίχνευσης που αποδίδουν εξαιρετικά σε ένα dataset αποτυγχάνουν πλήρως όταν αξιολογούνται σε άλλο, ακόμη και όταν το άλλο dataset περιέχει ομοειδείς τύπους επιθέσεων. Αυτό υποδεικνύει ότι τα περισσότερα μοντέλα «μαθαίνουν» τα dataset και όχι το underlying malicious behavior.

Οι λόγοι αυτής της αποτυχίας είναι πολυεπίεδοι. Πρώτον, διαφορετικά datasets καταγράφονται σε διαφορετικά περιβάλλοντα, με διαφορετική ένταση κίνησης, διαφορετικές συσκευές και διαφορετικές υπηρεσίες. Το IoT-23, για παράδειγμα, περιλαμβάνει traffic από πραγματικές consumer IoT συσκευές, ενώ το UNSW-NB15 δημιουργήθηκε χρησιμοποιώντας enterprise-level traffic generators. Τα δύο αυτά περιβάλλοντα παράγουν εντελώς διαφορετικά distributions, γεγονός που καθιστά αδύνατη τη μεταφορά γνώσης από το ένα μοντέλο στο άλλο χωρίς προσαρμογή.

Δεύτερον, τα datasets χρησιμοποιούν εντελώς διαφορετικά εργαλεία και μεθοδολογίες για την εξαγωγή ροών. Η ίδια κατηγορία επιθέσεων (π.χ. DoS) μπορεί να έχει διαφορετική υλοποίηση, διαφορετική διάρκεια, διαφορετική ένταση traffic και διαφορετικά statistical fingerprints. Τα μοντέλα ML, επομένως, μαθαίνουν patterns συγκεκριμένα του dataset καταγωγής και όχι της επίθεσης συμπεριφοράς καθ' αυτής.

Τρίτον, η ετερογένεια των χαρακτηριστικών επιδεινώνει περαιτέρω την κατάσταση. Όπως έχει αποδειχθεί στα E-XAI papers, η σχετική σημασία των χαρακτηριστικών μεταβάλλεται δραματικά μεταξύ datasets. Μάλιστα, μέσω SHAP analysis έχει φανεί ότι ορισμένα χαρακτηριστικά που είναι κρίσιμα στο CICIDS2017 (π.χ. «flow_duration») είναι σχεδόν άχρηστα στο UNSW-NB15, όπου το traffic είναι πιο σταθερό. Αυτή η ανακολουθία συνιστά dataset shift, το οποίο καθιστά τη γενίκευση σχεδόν αδύνατη χωρίς κατάλληλες τεχνικές domain adaptation.

Η έρευνα έχει αρχίσει να προτείνει λύσεις, αλλά βρίσκεται ακόμη σε εμβρυακό στάδιο. Το UM-NIDS (2024) και συναφή έργα προτείνουν unified preprocessing pipelines, όπου όλες οι ροές εξάγονται με την ίδια μεθοδολογία από raw PCAPs. Άλλες προσεγγίσεις αξιοποιούν adversarial domain adaptation, προσπαθώντας να ευθυγραμμίσουν τα distributions μεταξύ datasets. Το meta-learning έχει επίσης ενταχθεί στον χώρο, με στόχο τη δημιουργία μοντέλων που μαθαίνουν να προσαρμόζονται σε νέα δεδομένα με ελάχιστη επανεκπαίδευση.

Ωστόσο, μία από τις πλέον πρωτοποριακές ιδέες των τελευταίων ετών είναι η χρήση Explainable AI ως εργαλείο διάγνωσης και βελτίωσης της cross-dataset γενίκευσης. Με

XAI είναι δυνατόν να εντοπιστεί ποια χαρακτηριστικά διατηρούν σταθερή σημασία σε διαφορετικά περιβάλλοντα και ποια είναι dataset-specific artifacts. Με αυτό τον τρόπο, το XAI δε χρησιμοποιείται μόνο για ερμηνεία αποφάσεων, αλλά και ως instrument για τη βελτίωση της μεθοδολογικής εγκυρότητας των IDS. Η ενσωμάτωση αυτής της στρατηγικής αποτελεί σημαντική καινοτομία και ανοίγει δρόμους για ελαφριά, επεξηγήσιμα και πραγματικά generalizable IDS.

2.6 Ερευνητικά Κενά και Προοπτικές

Παρά την εκτεταμένη έρευνα στον χώρο της ανίχνευσης κυβερνοεπιθέσεων με τεχνικές μηχανικής μάθησης, η βιβλιογραφία αναδεικνύει σημαντικά κενά που περιορίζουν την αξιοπιστία, τη γενίκευση και την πρακτική εφαρμοσιμότητα των προτεινόμενων συστημάτων. Τα περισσότερα έργα εστιάζουν σε κλειστά ή μόνο-πηγαία datasets, σε στατικές offline αξιολογήσεις ή σε μεθοδολογίες που αγνοούν τις πραγματικές συνθήκες ετερογένειας της σύγχρονης δικτυακής κίνησης (Makridis,2023). Επιπλέον, η αξιοποίηση της Explainable AI (XAI) παραμένει κατακερματισμένη, με ελάχιστες ολοκληρωμένες προσεγγίσεις σε περιβάλλοντα flow-based ή cross-dataset αξιολόγησης (E-XAI Framework,2024).

Η ανάλυση των πηγών που παρατέθηκαν σε αυτή τη βιβλιογραφική ανασκόπηση οδηγεί στην ανάδειξη πέντε κύριων θεματικών κενών:

1. Έλλειψη cross-dataset generalization
2. Έλλειψη ρεαλιστικών ροών (flow-based, real-time) με υψηλής ποιότητας labels
3. Περιορισμένη ενσωμάτωση XAI σε ML μοντέλα IDS
4. Ασυνεπή preprocessing και μη αναπαραγώγιμες μεθοδολογίες
5. Μη επαρκής ανάλυση robustness και adversarial αντοχής των μοντέλων.

2.6.1 Περιορισμοί των Υφιστάμενων Datasets

Παρότι τα CICIDS2017 και UNSW-NB15 αποτελούν δύο από τα πιο δημοφιλή benchmarks, η βιβλιογραφία αναδεικνύει σημαντικές αδυναμίες τους. Για παράδειγμα, πρόσφατες μελέτες επιβεβαιώνουν ότι το CICIDS2017 περιλαμβάνει σφάλματα στα extracted flows, τα οποία διορθώθηκαν με νέα εργαλεία εξαγωγής χαρακτηριστικών, όπως το LycoSTand (LycosID2017), το οποίο βελτιώνει σημαντικά τις επιδόσεις ML μοντέλων (Strom et al., 2023). Αντίστοιχα, το UNSW-NB15 παρουσιάζει ετερογένεια

και μη ισορροπημένη κατανομή επιθέσεων, επηρεάζοντας την ακρίβεια και την ικανότητα γενίκευσης (Moustafa & Slay, 2015).

Ακόμη πιο σύγχρονα datasets, όπως το IoT-23 (Garcia et al., 2020), το BoT-IoT (Koroniotis et al., 2019) και το CICIoT2023 (Hindy et al., 2023), προσπαθούν να ενσωματώσουν ρεαλιστικό IoT traffic. Ωστόσο, παρουσιάζουν δύο βασικούς περιορισμούς:

- Έλλειψη cross-network heterogeneity
- Εστίαση σε συγκεκριμένα πρωτόκολλα ή επιθέσεις, κάτι που περιορίζει το generalizability

Οι παραπάνω αδυναμίες συνοψίζονται στον Πίνακα 1.

Πίνακας 1- Συγκριτική Ανάλυση Datasets

Dataset	Τύπος	Πλεονεκτήματα	Περιορισμοί	Πηγές
CICIDS2017	PCAP + flows	Πλούσιες επιθέσεις, Flow metadata	Flow extraction λάθη, overfitting	<i>Strom et al. (2023)</i>
UNSW-NB15	Realistic emulated traffic	Μεικτό περιβάλλον	Class imbalance, παλιό	<i>Moustafa & Slay (2015)</i>
IoT-23	IoT flows	Real devices	Περιορισμένο IoT diversity	<i>Garcia et al. (2020)</i>
BoT-IoT	IoT botnet	High attack volume	Extreme imbalance	<i>Koroniotis et al. (2019)</i>
CICIoT2023	IoT flows	105 IoT devices	Ακόμη νέο, όχι cross-site	<i>Hindy et al. (2023)</i>

2.6.2 Περιορισμοί Μεθοδολογίας και Απόδοσης ML Μοντέλων

Η πλειονότητα των μελετών επιτυγχάνει πολύ υψηλές επιδόσεις σε offline συνθήκες (accuracy 98%+), αλλά μόνο όταν εκπαιδεύεται και αξιολογείται στο ίδιο dataset. Σε

cross-dataset δοκιμές, τα ποσοστά πέφτουν δραματικά, ακόμη και κάτω από 40% (Hindy et al., 2023, Austin, 2021).

Πολλές μελέτες εφαρμόζουν CNNs, LSTM, BiLSTM, RF< XGBoost ή υβριδικές αρχιτεκτονικές, όπως PSO-LSTM (Rahman et al., 2024), επιτυγχάνοντας υψηλές επιδόσεις εντός του dataset. Ωστόσο, η βιβλιογραφία επιβεβαιώνει ότι:

- Τα DL μοντέλα «υπερεκπαιδεύονται» εύκολα
- Οι κλασσικοί classifiers χάνουν σε recall
- Τα hybrid μοντέλα είναι ακριβά σε inference latency

Συγκεντρωτικά, ο Πίνακας 2 αποτυπώνει τις εντοπισμένες επιδόσεις.

Πίνακας 2- Συνοπτική απόδοση ML μοντέλων

Μεθοδολογία	Accuracy	F1-score	Κύρια συμπεράσματα	Πηγές
Random Forest	95–99%	0.94–0.99	Καλή baseline, χαμηλή generalization	Strom et al. (2023)
LSTM / BiLSTM	98–99%	0.97–0.99	Overfitting, υψηλή latency	Rahman et al. (2024)
CNN-based IDS	97–99%	0.96–0.98	Φτωχή εξήγηση, static	Makridis (2023)
Hybrid PSO-LSTM	~99%	0.99	Υπολογιστικά βαρύ	Rahman et al. (2024)
XAI-augmented IDS	90–97%	–	Εξηγεί patterns αλλά μειώνει απόδοση	E-XAI (2024)

2.6.3 Ελλιπής Ενσωμάτωση Explainable AI

Παρά την ανάγκη για διαφάνεια, ειδικά σε συστήματα κυβερνοασφάλειας, η βιβλιογραφία δείχνει ότι:

- Τα περισσότερα XAI frameworks (π.χ. SHAP, LIME, Anchor) εφαρμόζονται μετά την ταξινόμηση
- Σχεδόν κανένα δεν ενσωματώνεται σε real-time flow-based IDS
- Ακόμη και τα πιο πρόσφατα frameworks, όπως το E-XAI (2024), αξιολογούνται σε black-box μοντέλα αλλά χωρίς real-world deployment

Ταυτόχρονα, η ανάλυση στο Industry 5.0 IDS (Makridis, 2023) δείχνει ότι τα XAI μοντέλα προσφέρουν:

- Εξηγήσιμη αποσαφήνιση προσβλητικής συμπεριφοράς
- Μείωση false positives
- Εντοπισμό zero-day patterns

Παρ' όλ' αυτά παραμένουν περιορισμένα σε computational cost consistency.

Το κενό είναι ξεκάθαρο:

Δεν υπάρχει επί του παρόντος σαφώς τεκμηριωμένη, cross-dataset, XAI-enhanced, flow-based μεθοδολογία ανίχνευσης ανωμαλιών. Αυτό αποτελεί σημαντική ευκαιρία για ερευνητική καινοτομία.

2.6.4 Προβλήματα Αναπαραγωγιμότητας και Ελλιπές Processing

Στις περισσότερες μελέτες, η processing διαδικασία δεν περιγράφεται επαρκώς, οδηγώντας σε :

- Μη αναπαραγώγιμες μεθοδολογίες
- Ασυνεπή data splits
- Αδιαφανείς μεθόδους feature selection
- Απώλεια temporal dependencies σε flows

Χαρακτηριστικό παράδειγμα είναι οι μελέτες που χρησιμοποιούν CICIDS2017, όπου η διαφορετική μέθοδος feature extraction οδηγεί σε αποκλίσεις στην τελική απόδοση έως και 30% (Strom et al., 2023).

2.6.5 Έλλειψη Robustness και Adversarial Testing

Ένα άλλο σοβαρό κενό είναι η περιορισμένη εξέταση robustness:

- Ελάχιστες εργασίες πραγματοποιούν adversarial attacks (FGSM, PGD)
- Καμία σχεδόν εργασία δεν εξετάζει adaptive attackers
- Τα DL μοντέλα αποδεικνύονται ευάλωτα ακόμη και σε μικρές perturbations

Αυτό υποδεικνύει ότι τα περισσότερα IDS συστήματα είναι έτοιμα για research benchmarking αλλά όχι για εφαρμογή σε πραγματικά δίκτυα.

Πίνακας 3- Research gaps που προκύπτουν

Θεματική Περιοχή	Research Gap	Πηγές
Generalization	IDS μοντέλα αποτυγχάνουν σε cross-dataset αξιολόγηση	<i>Hindy et al. (2023), Austin (2021)</i>
Flow-based ML	Έλλειψη real-time flow-based XAI	<i>E-XAI (2024), Makridis (2023)</i>
XAI	Περιορισμένη εφαρμογή σε IDS, χαμηλή integration	<i>E-XAI (2024)</i>
Robustness	Σχεδόν πλήρης απουσία adversarial testing	<i>Rahman et al. (2024)</i>
Preprocessing	Μη τυποποιημένα pipelines	<i>Strom et al. (2023)</i>

2.7 Σύνοψη

Από τη συνολική ανάλυση της βιβλιογραφίας προκύπτει ότι η αδυναμία γενίκευσης, η έλλειψη explainability και η ανάγκη για ρεαλιστική flow-based αξιολόγηση αποτελούν το τρίπτυχο στο οποίο συγκλίνουν όλες οι σύγχρονες ερευνητικές εργασίες. Συνεπώς, μία ερευνητική εργασία η οποία στοχεύει:

- Να συγκρίνει ML μοντέλα σε cross-dataset περιβάλλον
- Να εφαρμόσει flow-based ανάλυση και
- Να ενσωματώσει Explainable AI για διαφάνεια και ερμηνεία

Θα καλύψει ένα κρίσιμο κενό στη διεθνή βιβλιογραφία και θα προσφέρει πραγματική καινοτομία.

Η παρούσα διπλωματική εργασία λοιπόν, εστιάζει ακριβώς σε αυτή την κατεύθυνση.

3. Μεθοδολογία

Σκοπός του παρόντος κεφαλαίου είναι να παρουσιάσει αναλυτικά τη μεθοδολογική προσέγγιση που ακολουθήθηκε για την πειραματική αξιολόγηση μοντέλων Μηχανικής Μάθησης σε περιβάλλον ανίχνευσης εισβολών. Όπως αναδείχθηκε στα προηγούμενα κεφάλαια, ένα από τα σημαντικότερα προβλήματα των σύγχρονων ML-based Intrusion

Detection Systems είναι η περιορισμένη ικανότητα γενίκευσης όταν ένα μοντέλο εκπαιδεύεται σε ένα σύνολο δεδομένων και αξιολογείται σε ένα διαφορετικό. Το φαινόμενο αυτό, γνωστό ως cross-dataset generalization gap, αποτελεί κεντρικό ερευνητικό άξονα της παρούσας εργασίας.

Η πειραματική μεθοδολογία σχεδιάστηκε με στόχο να εξετάσει κατά πόσο τρία κλασικά μοντέλα supervised Machine Learning μπορούν να διατηρήσουν την απόδοσή τους όταν μεταφερθούν από ένα dataset σε άλλο. Για τον σκοπό αυτό χρησιμοποιήθηκαν δύο ευρέως διαδεδομένα datasets ανίχνευσης εισβολών, το CICIDS2017 και το UNSW-NB15. Η επιλογή τους δεν έγινε τυχαία, καθώς και τα δύο χρησιμοποιούνται συχνά στη διεθνή βιβλιογραφία, αλλά παρουσιάζουν σημαντικές διαφορές στη διαδικασία συλλογής, στο feature extraction, στην κατανομή των κλάσεων και στη φύση των επιθέσεων.

Η μεθοδολογία δεν στοχεύει στην ανάπτυξη ενός πλήρους real-time IDS έτοιμου για παραγωγική εγκατάσταση. Αντίθετα, επικεντρώνεται στην ελεγχόμενη πειραματική αξιολόγηση του κατά πόσο ένα μοντέλο που εμφανίζει υψηλή απόδοση εντός ενός dataset μπορεί πράγματι να μεταφέρει αυτή τη γνώση σε ένα διαφορετικό περιβάλλον. Επιπλέον, ενσωματώνεται ανάλυση Explainable AI μέσω SHAP, όχι ως μηχανισμός βελτίωσης της απόδοσης, αλλά ως εργαλείο ερμηνείας και ανάλυσης αποτυχίας των μοντέλων.

3.1 Ερευνητικός Σχεδιασμός

Η παρούσα εργασία ακολουθεί πειραματικό σχεδιασμό τεσσάρων βασικών σεναρίων. Τα δύο πρώτα σενάρια αποτελούν same-dataset baselines, όπου η εκπαίδευση και η αξιολόγηση πραγματοποιούνται σε διαφορετικά τμήματα του ίδιου dataset. Τα δύο επόμενα σενάρια αποτελούν cross-dataset πειράματα, όπου η εκπαίδευση γίνεται στο ένα dataset και η αξιολόγηση στο άλλο.

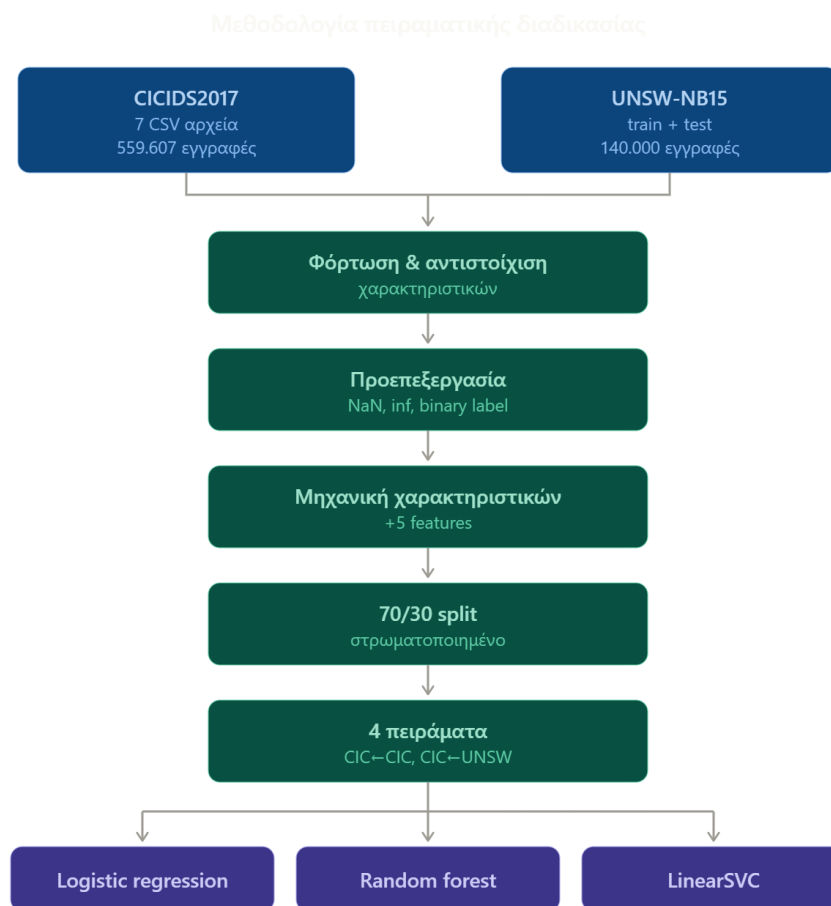
Τα πειράματα είναι τα εξής:

1. **CIC → CIC**: εκπαίδευση στο 70% του CICIDS2017 και αξιολόγηση στο υπόλοιπο 30%.
2. **UNSW → UNSW**: εκπαίδευση στο 70% του UNSW-NB15 και αξιολόγηση στο υπόλοιπο 30%.
3. **CIC → UNSW**: εκπαίδευση σε ολόκληρο το CICIDS2017 και αξιολόγηση σε ολόκληρο το UNSW-NB15.

4. **UNSW → CIC**: εκπαίδευση σε ολόκληρο το UNSW-NB15 και αξιολόγηση σε ολόκληρο το CICIDS2017.

Τα δύο πρώτα πειράματα χρησιμοποιούνται ως σημείο αναφοράς, ώστε να διαπιστωθεί η απόδοση των μοντέλων σε συνθήκες όπου η κατανομή εκπαίδευσης και δοκιμής προέρχεται από την ίδια πηγή. Τα δύο cross-dataset πειράματα εξετάζουν την πραγματική ικανότητα μεταφοράς γνώσης σε διαφορετικό feature distribution.

Η βασική ερευνητική υπόθεση είναι ότι τα μοντέλα θα εμφανίσουν υψηλή απόδοση στα same-dataset σενάρια, αλλά σημαντική υποβάθμιση στα cross-dataset σενάρια. Η υπόθεση αυτή ελέγχεται ποσοτικά μέσω μετρικών ταξινόμησης και ερμηνεύεται ποιοτικά μέσω SHAP και ανάλυσης μετατόπισης κατανομής χαρακτηριστικών.



Σχήμα 3.1: Ροή πειραματικής διαδικασίας.

3.2 Σύνολα Δεδομένων

3.2.1 CICIDS2017

Το CICIDS2017 αποτελεί ένα από τα πιο διαδεδομένα datasets για την αξιολόγηση συστημάτων ανίχνευσης εισβολών. Περιλαμβάνει δικτυακή κίνηση που έχει εξαχθεί σε μορφή flows μέσω του εργαλείου CICFlowMeter. Το dataset περιέχει τόσο benign traffic όσο και διαφορετικούς τύπους επιθέσεων, όπως DoS, DDoS, Port Scan, Brute Force, Web Attacks και άλλες μορφές κακόβουλης δραστηριότητας.

Στην παρούσα εργασία χρησιμοποιήθηκαν τα διαθέσιμα αρχεία τύπου WorkingHours.csv. Από αυτά επιλέχθηκαν οι στήλες που μπορούν να αντιστοιχιστούν σημασιολογικά με αντίστοιχες στήλες του UNSW-NB15. Μετά τη φόρτωση, τον καθαρισμό και την επιλογή χαρακτηριστικών, το τελικό σύνολο CICIDS2017 που χρησιμοποιήθηκε περιλάμβανε **559.607 εγγραφές**. Από αυτές, περίπου το 90,4% αντιστοιχούσε σε benign traffic και περίπου το 9,6% σε επιθέσεις.

Η δυαδική ετικέτα δημιουργήθηκε ως εξής: κάθε εγγραφή με τιμή BENIGN χαρακτηρίστηκε ως κλάση 0, ενώ κάθε διαφορετική τιμή στη στήλη Label χαρακτηρίστηκε ως κλάση 1. Επομένως, το πρόβλημα μετατράπηκε σε δυαδική ταξινόμηση:

- **0 = Benign**
- **1 = Attack**

Η συγκεκριμένη μετατροπή επιλέχθηκε ώστε η εργασία να εστιάσει στη γενική διάκριση μεταξύ κανονικής και κακόβουλης κίνησης, αντί για πολυταξική ταξινόμηση ανά τύπο επίθεσης. Η επιλογή αυτή μειώνει την πολυπλοκότητα και επιτρέπει πιο άμεση σύγκριση μεταξύ datasets με διαφορετικές κατηγορίες επιθέσεων.

3.2.2 UNSW-NB15

Το UNSW-NB15 αποτελεί επίσης ένα καθιερωμένο dataset στον χώρο των Network Intrusion Detection Systems. Περιλαμβάνει κανονική και κακόβουλη δικτυακή κίνηση, με διαφορετικές κατηγορίες επιθέσεων όπως Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode και Worms.

Σε αντίθεση με το CICIDS2017, το UNSW-NB15 παρέχεται ήδη σε αρχεία εκπαίδευσης και δοκιμής. Στην παρούσα εργασία τα διαθέσιμα train και test αρχεία ενώθηκαν σε ένα ενιαίο σύνολο, ώστε να εφαρμοστεί κοινή μεθοδολογία διαχωρισμού 70/30 στα same-

dataset πειράματα. Η τελική ομάδα δεδομένων που χρησιμοποιήθηκε περιλάμβανε **140.000 εγγραφές**.

Η στήλη label του UNSW-NB15 είναι ήδη δυαδική, επομένως δεν απαιτήθηκε μετατροπή από πολυταξική σε δυαδική μορφή. Η κλάση 0 αντιστοιχεί σε benign traffic, ενώ η κλάση 1 αντιστοιχεί σε attack traffic.

Σε αντίθεση με το CICIDS2017, το UNSW-NB15 παρουσιάζει πιο ισορροπημένη κατανομή μεταξύ benign και attack traffic. Συγκεκριμένα, στο σύνολο αξιολόγησης των 42.000 εγγραφών του same-dataset πειράματος, οι benign εγγραφές ήταν 22.005 και οι attack εγγραφές 19.995. Η διαφορά αυτή στην ισορροπία των κλάσεων αποτελεί σημαντικό στοιχείο της πειραματικής ανάλυσης, καθώς επηρεάζει την ερμηνεία μετρικών όπως accuracy, precision, recall και F1-score.

3.3 Επιλογή και Αντιστοίχιση Χαρακτηριστικών

Ένα από τα βασικά προβλήματα στη cross-dataset αξιολόγηση IDS μοντέλων είναι ότι διαφορετικά datasets χρησιμοποιούν διαφορετικά ονόματα, διαφορετικές μεθόδους εξαγωγής και διαφορετικές μονάδες μέτρησης χαρακτηριστικών. Το CICIDS2017 παράγεται μέσω CICFlowMeter, ενώ το UNSW-NB15 βασίζεται σε διαφορετική μεθοδολογία εξαγωγής χαρακτηριστικών. Για τον λόγο αυτό δεν είναι δυνατή η απευθείας χρήση όλων των διαθέσιμων στηλών.

Η εργασία περιορίστηκε σε ένα κοινό σύνολο χαρακτηριστικών που μπορούν να αντιστοιχιστούν σημασιολογικά μεταξύ των δύο datasets. Τα βασικά χαρακτηριστικά που χρησιμοποιήθηκαν είναι τα ακόλουθα:

Κοινό χαρακτηριστικό	CICIDS2017	UNSW- NB15	Περιγραφή
Dur	Flow Duration	dur	Διάρκεια ροής
Spkts	Total Fwd Packets	spkts	Πακέτα από την πηγή
Dpkts	Total Backward Packets	dpkts	Πακέτα προς την πηγή/προορισμό
Sbytes	Total Length of Fwd Packets	sbytes	Bytes από την πηγή

Κοινό χαρακτηριστικό	CICIDS2017	UNSW- NB15	Περιγραφή
Dbytes	Total Length of Bwd Packets	dbytes	Bytes προς την πηγή/προορισμό
Rate	Flow Packets/s	rate	Ρυθμός πακέτων ανά δευτερόλεπτο

Η επιλογή περιορισμένου feature set έγινε για λόγους συγκρισιμότητας. Αν χρησιμοποιούνταν όλα τα χαρακτηριστικά κάθε dataset, η cross-dataset αξιολόγηση θα ήταν μεθοδολογικά προβληματική, επειδή τα μοντέλα θα εκπαιδεύονταν σε διαφορετικό feature space από αυτό στο οποίο θα αξιολογούνταν.

Πρέπει να σημειωθεί ότι η αντιστοίχιση αυτή είναι σημασιολογική και όχι πλήρης εναρμόνιση μονάδων. Για παράδειγμα, η διάρκεια ροής στο CICIDS2017 καταγράφεται σε μικροδευτερόλεπτα, ενώ στο UNSW-NB15 σε δευτερόλεπτα. Η διαφορά αυτή δεν διορθώθηκε χειροκίνητα στα κύρια πειράματα, καθώς αποτελεί μέρος του προβλήματος feature distribution shift που εξετάζεται στην εργασία. Η επιλογή αυτή επιτρέπει να φανεί πώς ακόμη και φαινομενικά αντίστοιχα χαρακτηριστικά μπορούν να οδηγήσουν σε αποτυχία γενίκευσης όταν προέρχονται από διαφορετικά εργαλεία και διαφορετικές διαδικασίες εξαγωγής.

3.4 Μηχανική Χαρακτηριστικών

Πέρα από τα έξι βασικά χαρακτηριστικά, δημιουργήθηκαν πέντε επιπλέον engineered features. Τα χαρακτηριστικά αυτά υπολογίστηκαν αποκλειστικά από τις τιμές κάθε εγγραφής, χωρίς χρήση πληροφορίας από άλλες γραμμές του dataset. Με αυτόν τον τρόπο αποφεύχθηκε η διαρροή πληροφορίας μεταξύ train και test.

Τα engineered features είναι τα εξής:

1. **pkt_ratio** = **spkts** / (**dpkts** + 1)
Εκφράζει την ασυμμετρία μεταξύ πακέτων πηγής και προορισμού.
2. **byte_ratio** = **sbytes** / (**dbytes** + 1)
Εκφράζει την ασυμμετρία μεταξύ bytes πηγής και προορισμού.
3. **bytes_per_packet** = **sbytes** / (**spkts** + 1)
Υπολογίζει τον μέσο αριθμό bytes ανά πακέτο από την πλευρά της πηγής.

4. **log_rate** = **log1p(rate)**

Λογαριθμικός μετασχηματισμός του ρυθμού πακέτων, ώστε να μειωθεί η επίδραση ακραίων τιμών.

5. **log_dur** = **log1p(dur)**

Λογαριθμικός μετασχηματισμός της διάρκειας ροής, ώστε να συμπιεστεί η έντονα ασύμμετρη κατανομή.

Η προσθήκη του +1 στους παρονομαστές εφαρμόστηκε για την αποφυγή διαίρεσης με το μηδέν. Η συνάρτηση log1p χρησιμοποιήθηκε αντί του απλού λογαρίθμου, επειδή είναι κατάλληλη για περιπτώσεις όπου οι τιμές μπορεί να είναι μηδενικές ή πολύ μικρές. Συνολικά, κάθε εγγραφή αναπαρίσταται από **11 χαρακτηριστικά**: έξι βασικά και πέντε engineered.

3.5 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία εφαρμόστηκε με στόχο την αποφυγή σφαλμάτων, τη διατήρηση της αναπαραγωγιμότητας και την αποτροπή data leakage. Η διαδικασία περιλαμβάνει τα εξής στάδια:

1. Φόρτωση των αρχείων CSV.
2. Καθαρισμό των ονομάτων στηλών.
3. Αντιστοίχιση των στηλών CICIDS2017 στα κοινά ονόματα χαρακτηριστικών.
4. Μετατροπή αριθμητικών στηλών σε αριθμητική μορφή.
5. Αντικατάσταση τιμών +inf και -inf με NaN.
6. Αφαίρεση εγγραφών με ελλιπείς τιμές στα επιλεγμένα χαρακτηριστικά ή στην ετικέτα.
7. Δημιουργία engineered features.
8. Νέα αφαίρεση εγγραφών με μη έγκυρες τιμές μετά τη μηχανική χαρακτηριστικών.
9. Διαχωρισμό των δεδομένων σε train και test όπου απαιτείται.
10. Κανονικοποίηση μέσω RobustScaler εντός του pipeline κάθε μοντέλου.

Η κανονικοποίηση πραγματοποιήθηκε με **RobustScaler**. Ο RobustScaler χρησιμοποιεί τη διάμεσο και το ενδοτεταρτημοριακό εύρος, αντί για τη μέση τιμή και την τυπική απόκλιση. Αυτό τον καθιστά πιο ανθεκτικό σε ακραίες τιμές, οι οποίες εμφανίζονται συχνά σε δικτυακά δεδομένα.

Ιδιαίτερη προσοχή δόθηκε στο να μην υπάρξει διαρροή πληροφορίας από το test set προς το train set. Για τον λόγο αυτό, ο scaler εκπαιδεύεται μόνο στα δεδομένα εκπαίδευσης και στη συνέχεια εφαρμόζεται στα δεδομένα αξιολόγησης. Στα cross-dataset πειράματα, αυτό σημαίνει ότι ο scaler μαθαίνει τα στατιστικά του source dataset και εφαρμόζεται αυτούσιος στο target dataset. Αυτή η επιλογή είναι σημαντική, διότι προσομοιώνει ένα πραγματικό σενάριο μεταφοράς μοντέλου σε νέο περιβάλλον, όπου οι στατιστικές του νέου περιβάλλοντος δεν είναι διαθέσιμες κατά την εκπαίδευση.

3.6 Αντιμετώπιση Ανισορροπίας Κλάσεων

Τα δύο datasets εμφανίζουν διαφορετικά επίπεδα ανισορροπίας. Το CICIDS2017 είναι έντονα ανισόρροπο, καθώς η benign κίνηση αποτελεί τη μεγάλη πλειονότητα των εγγραφών. Αντίθετα, το UNSW-NB15 είναι πιο ισορροπημένο, με πολύ μεγαλύτερο ποσοστό attack traffic.

Για την αντιμετώπιση της ανισορροπίας δεν εφαρμόστηκε oversampling ή undersampling. Η επιλογή αυτή έγινε για να μη μεταβληθεί τεχνητά η κατανομή των δεδομένων και για να διατηρηθούν όλες οι εγγραφές. Αντί για δειγματοληπτική παρέμβαση, χρησιμοποιήθηκαν class weights στα μοντέλα.

Στη Logistic Regression και στο LinearSVC χρησιμοποιήθηκε η παράμετρος:

```
class_weight = "balanced"
```

Στο Random Forest χρησιμοποιήθηκε η παράμετρος:

```
class_weight = "balanced_subsample"
```

Με αυτόν τον τρόπο, οι εγγραφές της μειονοτικής κλάσης λαμβάνουν μεγαλύτερη βαρύτητα κατά την εκπαίδευση, χωρίς να αλλοιώνεται το αρχικό dataset.

3.7 Μοντέλα Μηχανικής Μάθησης

Στην πειραματική διαδικασία αξιολογήθηκαν τρία supervised Machine Learning μοντέλα:

1. Logistic Regression
2. Random Forest
3. Linear Support Vector Classifier

Η επιλογή των συγκεκριμένων μοντέλων έγινε επειδή αντιπροσωπεύουν διαφορετικές κατηγορίες ταξινομητών και επιτρέπουν σύγκριση μεταξύ γραμμικών και μη γραμμικών προσεγγίσεων.

3.7.1 Logistic Regression

Η Logistic Regression χρησιμοποιήθηκε ως γραμμικό baseline μοντέλο. Είναι απλή, ερμηνεύσιμη και κατάλληλη για δυαδικά προβλήματα ταξινόμησης. Το μοντέλο ρυθμίστηκε με `max_iter=5000` και `class_weight="balanced"`.

Επειδή η Logistic Regression παράγει πιθανότητες πρόβλεψης, εφαρμόστηκε βελτιστοποίηση του decision threshold. Αντί να χρησιμοποιηθεί αυτόματα το κλασικό όριο 0,50, πραγματοποιήθηκε αναζήτηση σε τιμές από 0,30 έως 0,60 με βήμα 0,05. Η επιλογή του threshold έγινε αποκλειστικά πάνω στα δεδομένα εκπαίδευσης, με βάση τη βελτιστοποίηση του F1-score. Έτσι αποφεύχθηκε η χρήση πληροφορίας από το test set.

3.7.2 Random Forest

Το Random Forest επιλέχθηκε ως μη γραμμικό ensemble μοντέλο. Είναι ιδιαίτερα δημοφιλές σε προβλήματα tabular data και συχνά επιτυγχάνει υψηλή απόδοση σε IDS datasets. Το μοντέλο διαμορφώθηκε με 120 δέντρα, μέγιστο βάθος 12, `min_samples_leaf=5`, `min_samples_split=10` και `class_weight="balanced_subsample"`.

Οι περιορισμοί στο βάθος και στα ελάχιστα δείγματα ανά φύλλο εφαρμόστηκαν για να περιοριστεί η υπερπροσαρμογή. Παρόλα αυτά, ένα από τα ζητήματα που εξετάζεται στην εργασία είναι κατά πόσο ένα ισχυρό ensemble μοντέλο, όπως το Random Forest, μαθαίνει γενικεύσιμα μοτίβα επίθεσης ή απλώς dataset-specific patterns.

3.7.3 LinearSVC

Το LinearSVC χρησιμοποιήθηκε ως δεύτερο γραμμικό μοντέλο, βασισμένο στην αρχή μεγιστοποίησης του περιθωρίου μεταξύ των κλάσεων. Ρυθμίστηκε με `class_weight="balanced"` και `max_iter=10000`.

Σε αντίθεση με τη Logistic Regression, το LinearSVC δεν παράγει άμεσα πιθανότητες πρόβλεψης. Για τον λόγο αυτό δεν εφαρμόστηκε threshold tuning. Το μοντέλο χρησιμοποιήθηκε ως έχει, με το προεπιλεγμένο όριο απόφασης του ταξινομητή.

3.8 Σχεδιασμός Πειραμάτων

Ο πειραματικός σχεδιασμός βασίστηκε σε τέσσερα σενάρια αξιολόγησης.

Στα same-dataset πειράματα, κάθε dataset χωρίστηκε σε 70% train και 30% test με stratified split. Η διαστρωμάτωση εφαρμόστηκε ώστε η αναλογία benign και attack εγγραφών να παραμείνει σταθερή μεταξύ train και test. Για λόγους αναπαραγωγιμότητας χρησιμοποιήθηκε σταθερό seed ίσο με 42.

Στο CIC → CIC πείραμα χρησιμοποιήθηκαν 391.724 εγγραφές για εκπαίδευση και 167.883 για αξιολόγηση. Στο UNSW → UNSW πείραμα χρησιμοποιήθηκαν 98.000 εγγραφές για εκπαίδευση και 42.000 για αξιολόγηση.

Στα cross-dataset πειράματα, το μοντέλο εκπαιδεύτηκε σε ολόκληρο το source dataset και αξιολογήθηκε σε ολόκληρο το target dataset. Συγκεκριμένα, στο CIC → UNSW το μοντέλο εκπαιδεύτηκε στις 559.607 εγγραφές του CICIDS2017 και αξιολογήθηκε στις 140.000 εγγραφές του UNSW-NB15. Αντίστοιχα, στο UNSW → CIC το μοντέλο εκπαιδεύτηκε στις 140.000 εγγραφές του UNSW-NB15 και αξιολογήθηκε στις 559.607 εγγραφές του CICIDS2017.

Η χρήση ολόκληρου του source dataset στα cross-dataset σενάρια έγινε επειδή το target dataset είναι πλήρως ανεξάρτητο από τα δεδομένα εκπαίδευσης. Με αυτόν τον τρόπο, το πείραμα εξετάζει κατά πόσο η μέγιστη διαθέσιμη πληροφορία από ένα dataset μπορεί να μεταφερθεί σε διαφορετικό περιβάλλον.

3.9 Μετρικές Αξιολόγησης

Για την αξιολόγηση των μοντέλων χρησιμοποιήθηκαν οι ακόλουθες μετρικές:

- **Accuracy:** το συνολικό ποσοστό σωστών προβλέψεων.
- **Balanced Accuracy:** ο μέσος όρος recall ανά κλάση.
- **Precision:** το ποσοστό των προβλέψεων attack που ήταν πράγματι attack.
- **Recall:** το ποσοστό των πραγματικών attacks που ανιχνεύθηκαν σωστά.
- **F1-score:** ο αρμονικός μέσος precision και recall.
- **Confusion Matrix:** ανάλυση των TN, FP, FN και TP.
- **Delta F1 (δF1):** διαφορά μεταξύ same-dataset F1 και cross-dataset F1 για το ίδιο μοντέλο και source dataset.

Το accuracy δεν θεωρήθηκε από μόνο του επαρκής μετρική, λόγω της έντονης ανισορροπίας στο CICIDS2017. Ένα μοντέλο που προβλέπει σχεδόν όλες τις εγγραφές ως benign μπορεί να επιτύχει υψηλό accuracy, χωρίς να έχει ουσιαστική ικανότητα ανίχνευσης επιθέσεων. Για τον λόγο αυτό, δόθηκε ιδιαίτερη έμφαση στο F1-score, στο recall και στο balanced accuracy.

Το Delta F1 χρησιμοποιήθηκε ειδικά για να αποτυπωθεί η πτώση απόδοσης από τα same-dataset στα cross-dataset σενάρια. Ορίζεται ως:

$$\delta F1 = F1_same-dataset - F1_cross-dataset$$

Μεγάλη τιμή $\delta F1$ υποδηλώνει μεγάλη απώλεια γενίκευσης.

3.10 Ανάλυση Μετατόπισης Κατανομής Χαρακτηριστικών

Για να ερμηνευθεί η πιθανή αποτυχία των μοντέλων στα cross-dataset σενάρια, πραγματοποιήθηκε ανάλυση της μετατόπισης κατανομής χαρακτηριστικών μεταξύ CICIDS2017 και UNSW-NB15. Για κάθε ένα από τα 11 χαρακτηριστικά υπολογίστηκαν:

- μέση τιμή,
- διάμεσος,
- ενδοτεταρτημοριακό εύρος,
- απόλυτη διαφορά διαμέσων,
- απόλυτη διαφορά IQR.

Η ανάλυση αυτή είναι αλγοριθμικά ανεξάρτητη και επιτρέπει να εξεταστεί αν οι διαφορές στην απόδοση οφείλονται μόνο στα μοντέλα ή και στη δομική ασυμβατότητα των δεδομένων. Ιδιαίτερο ενδιαφέρον παρουσιάζουν χαρακτηριστικά όπως το dur και το rate, τα οποία εμφανίζουν μεγάλες αποκλίσεις μεταξύ των δύο datasets. Η διαφορά αυτή σχετίζεται τόσο με διαφορετικές μονάδες μέτρησης όσο και με διαφορετικές διαδικασίες εξαγωγής χαρακτηριστικών.

Η συγκεκριμένη ανάλυση συνδέεται άμεσα με το ερευνητικό ερώτημα της εργασίας, καθώς δείχνει ότι η αποτυχία cross-dataset γενίκευσης δεν είναι απαραίτητα αποτέλεσμα αδυναμίας ενός συγκεκριμένου αλγορίθμου, αλλά μπορεί να πηγάζει από βαθύτερη ασυμβατότητα του feature space.

3.11 Explainable Failure Analysis με SHAP

Για την ερμηνεία των αποφάσεων του Random Forest χρησιμοποιήθηκε η μέθοδος SHAP. Η επιλογή του SHAP έγινε επειδή παρέχει ποσοτική εκτίμηση της συμβολής κάθε χαρακτηριστικού στην πρόβλεψη του μοντέλου. Ειδικά για το Random Forest χρησιμοποιήθηκε TreeExplainer, ο οποίος είναι κατάλληλος για μοντέλα βασισμένα σε δέντρα.

Η SHAP ανάλυση εφαρμόστηκε μόνο στο Random Forest για δύο λόγους. Πρώτον, το Random Forest ήταν το ισχυρότερο μοντέλο στα same-dataset πειράματα, επομένως είχε ιδιαίτερο ενδιαφέρον να εξεταστεί γιατί αποτυγχάνει όταν μεταφέρεται σε άλλο dataset. Δεύτερον, το TreeExplainer επιτρέπει αποδοτικό υπολογισμό SHAP values για tree-based μοντέλα.

Για λόγους υπολογιστικής αποδοτικότητας, η SHAP ανάλυση εφαρμόστηκε σε τυχαίο δείγμα έως 5.000 εγγραφών από το test set κάθε πειράματος. Η επιλογή του δείγματος έγινε με σταθερό seed, ώστε τα αποτελέσματα να είναι αναπαραγώγιμα.

Η SHAP ανάλυση δεν χρησιμοποιήθηκε ως απόδειξη επιτυχούς ανίχνευσης. Αντίθετα, χρησιμοποιήθηκε ως εργαλείο failure analysis. Στα cross-dataset σενάρια, όπου τα μοντέλα παρουσιάζουν έντονη πτώση απόδοσης, τα SHAP values βοηθούν να εξεταστεί ποια χαρακτηριστικά εξακολουθεί να θεωρεί σημαντικά το μοντέλο και κατά πόσο αυτά τα χαρακτηριστικά παραμένουν σταθερά μεταξύ datasets.

Με αυτόν τον τρόπο, η XAI ανάλυση συνδέεται άμεσα με το βασικό ερευνητικό πρόβλημα της εργασίας: εάν τα σημαντικά χαρακτηριστικά αλλάζουν ριζικά μεταξύ datasets, τότε η υψηλή απόδοση σε ένα dataset δεν αρκεί για να θεωρηθεί ότι το μοντέλο έχει μάθει γενικεύσιμα μοτίβα επιθέσεων.

3.12 Υλοποίηση και Αναπαραγωγικότητα

Η πειραματική διαδικασία υλοποιήθηκε σε Python, με χρήση βιβλιοθηκών όπως pandas, NumPy, scikit-learn, Matplotlib και SHAP. Όλα τα μοντέλα ενσωματώθηκαν σε pipelines του scikit-learn, όπου το πρώτο βήμα ήταν ο RobustScaler και το δεύτερο το αντίστοιχο μοντέλο ταξινόμησης. Η χρήση pipeline διασφαλίζει ότι η κανονικοποίηση εφαρμόζεται με συνέπεια και χωρίς διαρροή δεδομένων.

Για την αναπαραγωγιμότητα των αποτελεσμάτων χρησιμοποιήθηκε σταθερό seed ίσο με 42 σε όλα τα στάδια όπου υπήρχε τυχαιότητα: train-test split, Random Forest και επιλογή δείγματος για SHAP analysis.

Τα αποτελέσματα κάθε πειράματος αποθηκεύτηκαν σε αρχεία CSV, ενώ τα confusion matrices και τα SHAP plots αποθηκεύτηκαν ως εικόνες. Επιπλέον, δημιουργήθηκαν συνοπτικοί πίνακες με τις βασικές μετρικές αξιολόγησης, ώστε να διευκολυνθεί η σύγκριση μεταξύ μοντέλων και πειραματικών σεναρίων.

3.13 Περιορισμοί της Μεθοδολογίας

Η παρούσα μεθοδολογία έχει ορισμένους περιορισμούς, οι οποίοι πρέπει να ληφθούν υπόψη κατά την ερμηνεία των αποτελεσμάτων.

- Πρώτον, η εργασία περιορίζεται σε δυαδική ταξινόμηση benign/attack και δεν εξετάζει πολυταξική ταξινόμηση ανά τύπο επίθεσης. Η επιλογή αυτή έγινε για να επιτευχθεί συγκρισιμότητα μεταξύ datasets με διαφορετικές κατηγορίες επιθέσεων.
- Δεύτερον, χρησιμοποιείται περιορισμένο κοινό feature set 11 χαρακτηριστικών. Αυτό επιτρέπει cross-dataset αξιολόγηση, αλλά αποκλείει χαρακτηριστικά που ενδέχεται να είναι χρήσιμα εντός κάθε dataset ξεχωριστά.
- Τρίτον, δεν αξιολογούνται μοντέλα Deep Learning. Η επιλογή αυτή είναι σκόπιμη, καθώς ο στόχος της εργασίας δεν είναι η μέγιστη δυνατή απόδοση εντός ενός dataset, αλλά η καθαρή αξιολόγηση της γενίκευσης σε διαφορετικά feature distributions.
- Τέταρτον, η XAI ανάλυση βασίζεται στο SHAP και εφαρμόζεται στο Random Forest. Δεν πραγματοποιείται συγκριτική αξιολόγηση πολλαπλών XAI frameworks, όπως LIME ή Integrated Gradients. Το SHAP χρησιμοποιείται ως εργαλείο ερμηνείας και ανάλυσης αποτυχίας, όχι ως μηχανισμός βελτιστοποίησης του μοντέλου.
- Πέμπτον, η σημασιολογική αντιστοίχιση χαρακτηριστικών δεν συνεπάγεται πλήρη εναρμόνιση μονάδων μέτρησης. Αυτό αποτελεί ταυτόχρονα περιορισμό και αντικείμενο ανάλυσης, καθώς αναδεικνύει ένα κρίσιμο πρόβλημα στη χρήση δημόσιων IDS datasets για cross-dataset αξιολόγηση.

3.14 Σύνοψη Κεφαλαίου

Στο παρόν κεφάλαιο παρουσιάστηκε η πειραματική μεθοδολογία που ακολουθήθηκε για την αξιολόγηση της cross-dataset γενίκευσης μοντέλων Μηχανικής Μάθησης σε δεδομένα ανίχνευσης εισβολών. Χρησιμοποιήθηκαν τα datasets CICIDS2017 και UNSW-NB15, τα οποία μετατράπηκαν σε κοινό δυαδικό πρόβλημα ταξινόμησης benign/attack. Επιλέχθηκε κοινό feature space 11 χαρακτηριστικών, αποτελούμενο από έξι βασικά και πέντε engineered features.

Αξιολογήθηκαν τρία μοντέλα supervised Machine Learning: Logistic Regression, Random Forest και LinearSVC. Τα μοντέλα εξετάστηκαν τόσο σε same-dataset όσο και σε cross-dataset σενάρια. Η αξιολόγηση πραγματοποιήθηκε με accuracy, balanced accuracy, precision, recall, F1-score, confusion matrices και Delta F1. Επιπλέον, εφαρμόστηκε SHAP analysis στο Random Forest, με στόχο την ερμηνεία των αποφάσεων και την ανάλυση της αποτυχίας γενίκευσης.

Η μεθοδολογία αυτή επιτρέπει την εμπειρική διερεύνηση του βασικού ερευνητικού προβλήματος της εργασίας: κατά πόσο τα ML-based IDS μοντέλα μαθαίνουν γενικεύσιμα μοτίβα κακόβουλης κίνησης ή απλώς προσαρμόζονται στις ιδιαιτερότητες του dataset στο οποίο εκπαιδεύονται.

4. Πειραματικά Αποτελέσματα & Ανάλυση

Στο παρόν κεφάλαιο παρουσιάζονται και αναλύονται τα πειραματικά αποτελέσματα της εργασίας. Όπως περιγράφηκε στο Κεφάλαιο 4, η πειραματική διαδικασία βασίστηκε σε δύο δημόσια διαθέσιμα datasets ανίχνευσης εισβολών, το CICIDS2017 και το UNSW-NB15. Τα δύο datasets μετατράπηκαν σε κοινό δυαδικό πρόβλημα ταξινόμησης, όπου η κλάση 0 αντιστοιχεί σε κανονική δικτυακή κίνηση, ενώ η κλάση 1 αντιστοιχεί σε κακόβουλη δραστηριότητα.

Η αξιολόγηση πραγματοποιήθηκε με τρία μοντέλα supervised Machine Learning: Logistic Regression, Random Forest και LinearSVC. Τα μοντέλα εξετάστηκαν σε τέσσερα πειραματικά σενάρια. Τα δύο πρώτα σενάρια αποτελούν same-dataset αξιολόγηση, όπου το μοντέλο εκπαιδεύεται και αξιολογείται σε διαφορετικά τμήματα του ίδιου dataset. Τα δύο επόμενα σενάρια αποτελούν cross-dataset αξιολόγηση, όπου το μοντέλο εκπαιδεύεται στο ένα dataset και αξιολογείται στο άλλο.

Ο βασικός στόχος του κεφαλαίου είναι να εξεταστεί αν η υψηλή απόδοση που εμφανίζουν τα μοντέλα εντός ενός dataset μπορεί να διατηρηθεί όταν το μοντέλο εφαρμόζεται σε διαφορετική κατανομή δεδομένων. Η ανάλυση επικεντρώνεται κυρίως στο F1-score, στο recall, στο precision, στο balanced accuracy και στα confusion matrices, καθώς το απλό accuracy μπορεί να είναι παραπλανητικό σε ανισόρροπα IDS datasets.

4.1 Συνοπτική Παρουσίαση Αποτελεσμάτων

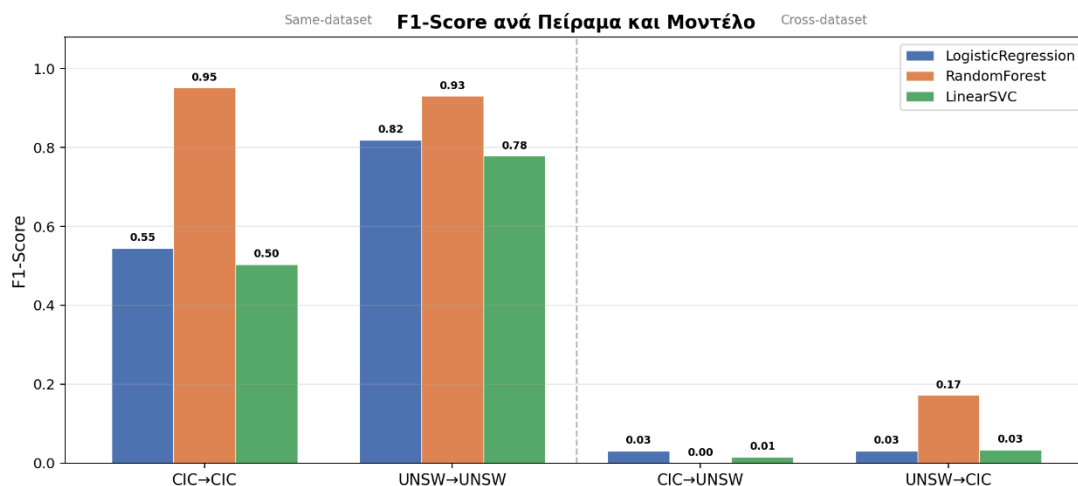
Ο Πίνακας 4 παρουσιάζει τα συνολικά αποτελέσματα όλων των πειραμάτων. Οι τιμές εμφανίζονται σε ποσοστιαία μορφή για ευκολότερη σύγκριση.

Πίνακας 4: Συνοπτικά αποτελέσματα πειραμάτων

Πείραμα	Μοντέλο	Accuracy	Balanced Accuracy	Precision	Recall	F1-score	δF1
CIC→CIC	Logistic Regression	86,3%	86,0%	40,0%	85,7%	54,5%	0,000
CIC→CIC	Random Forest	99,0%	99,2%	91,3%	99,3%	95,2%	0,000
CIC→CIC	LinearSVC	83,8%	84,7%	35,7%	85,8%	50,4%	0,000
UNSW→ UNSW	Logistic Regression	80,5%	81,1%	73,3%	92,8%	81,9%	0,000
UNSW→ UNSW	Random Forest	93,2%	93,4%	90,3%	96,0%	93,1%	0,000
UNSW→ UNSW	LinearSVC	80,5%	80,1%	84,6%	72,2%	77,9%	0,000
CIC→UNSW	Logistic Regression	53,1%	50,7%	91,4%	1,6%	3,1%	0,515
CIC→UNSW	Random Forest	52,4%	50,0%	0,0%	0,0%	0,0%	0,952
CIC→UNSW	LinearSVC	52,7%	50,4%	95,3%	0,8%	1,5%	0,489

Πείραμα	Μοντέλο	Accuracy	Balanced Accuracy	Precision	Recall	F1-score	δF1
UNSW→CIC	Logistic Regression	82,9%	47,1%	3,4%	2,9%	3,1%	0,788
UNSW→CIC	Random Forest	38,6%	51,0%	9,9%	66,2%	17,2%	0,759
UNSW→CIC	LinearSVC	82,2%	46,9%	3,6%	3,3%	3,4%	0,745

Από τον πίνακα προκύπτει άμεσα η βασική τάση της εργασίας. Στα same-dataset πειράματα, τα μοντέλα εμφανίζουν από μέτρια έως πολύ υψηλή απόδοση. Αντίθετα, στα cross-dataset πειράματα, η απόδοση μειώνεται δραματικά. Το φαινόμενο αυτό επιβεβαιώνει ότι η επιτυχία ενός μοντέλου σε ένα dataset δεν συνεπάγεται απαραίτητα ικανότητα γενίκευσης σε διαφορετικό dataset.



Σχήμα 4.1: Σύγκριση F1-score ανά πείραμα και μοντέλο.

4.2 Same-Dataset Αποτελέσματα

4.2.1 Πείραμα CIC→CIC

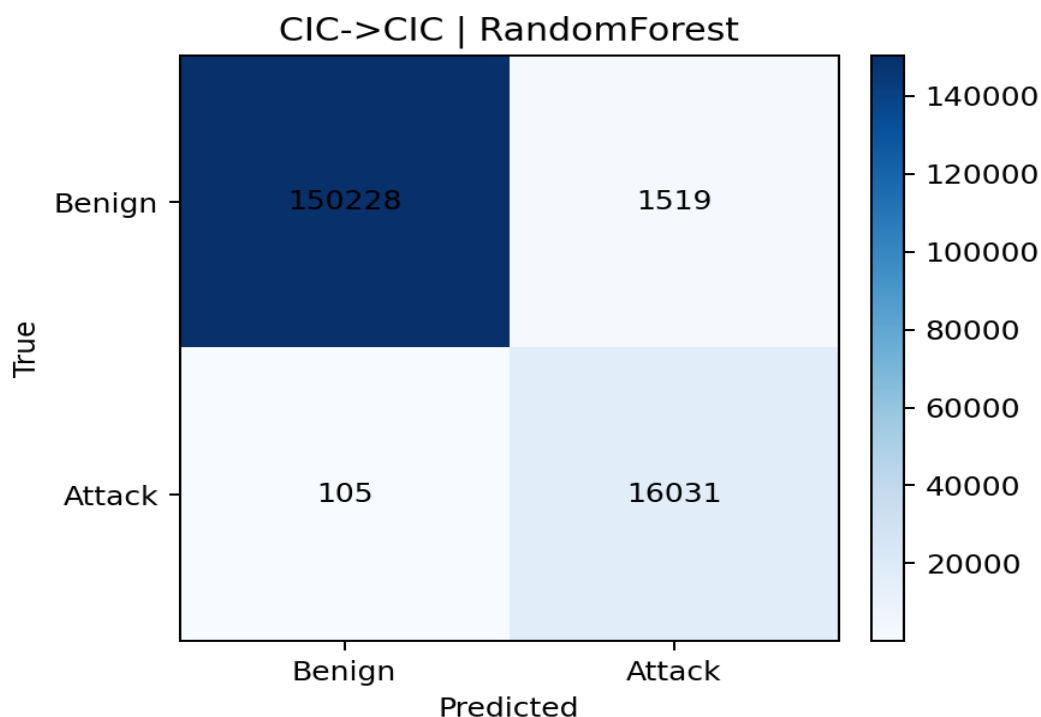
Στο πρώτο πείραμα, τα μοντέλα εκπαιδεύτηκαν στο 70% του CICIDS2017 και αξιολογήθηκαν στο υπόλοιπο 30%. Το πείραμα αυτό λειτουργεί ως baseline για το

CICIDS2017, καθώς η κατανομή εκπαίδευσης και αξιολόγησης προέρχεται από την ίδια πηγή.

Το Random Forest παρουσίασε την καλύτερη συνολική απόδοση, με accuracy 99,0%, balanced accuracy 99,2% και F1-score 95,2%. Το recall έφτασε το 99,3%, γεγονός που δείχνει ότι το μοντέλο εντόπισε σχεδόν όλες τις πραγματικές επιθέσεις. Παράλληλα, το precision 91,3% δείχνει ότι οι περισσότερες προβλέψεις attack ήταν σωστές.

Η Logistic Regression και το LinearSVC εμφάνισαν χαμηλότερη απόδοση. Και τα δύο γραμμικά μοντέλα πέτυχαν αρκετά υψηλό recall, περίπου 85–86%, αλλά είχαν σημαντικά χαμηλότερο precision. Η Logistic Regression είχε precision 40,0% και F1-score 54,5%, ενώ το LinearSVC είχε precision 35,7% και F1-score 50,4%. Αυτό σημαίνει ότι τα γραμμικά μοντέλα κατάφεραν να εντοπίσουν μεγάλο μέρος των επιθέσεων, αλλά παρήγαγαν πολλούς ψευδείς συναγερμούς.

Η διαφορά αυτή δείχνει ότι το Random Forest αξιοποιεί καλύτερα τις μη γραμμικές σχέσεις των χαρακτηριστικών στο CICIDS2017. Ωστόσο, όπως θα φανεί στη συνέχεια, αυτή η υψηλή απόδοση εντός του ίδιου dataset δεν μεταφέρεται απαραίτητα σε διαφορετικό dataset.



Σχήμα 2: Confusion Matrix του Random Forest στο CIC→CIC

Στο confusion matrix του Random Forest για το CIC→CIC παρατηρούνται τα εξής:

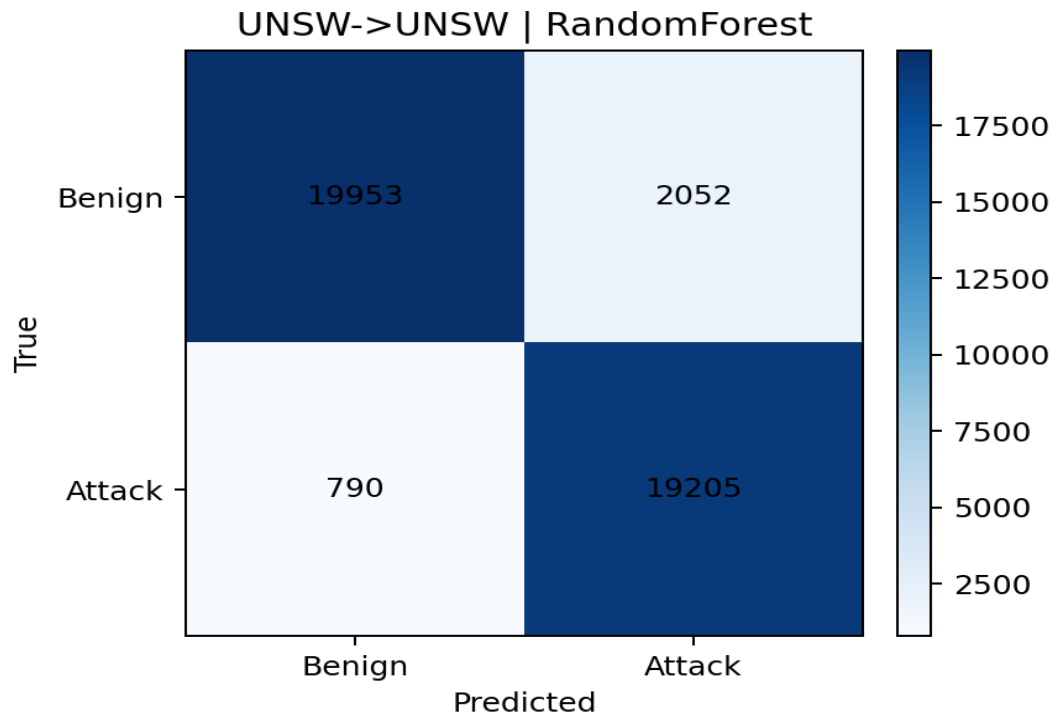
- True Negatives: 150.228
- False Positives: 1.519
- False Negatives: 105
- True Positives: 16.031

Το μοντέλο ταξινομήσε σωστά τη συντριπτική πλειονότητα των benign και attack εγγραφών. Τα μόλις 105 false negatives δείχνουν ότι το μοντέλο έχασε ελάχιστες επιθέσεις. Παράλληλα, τα 1.519 false positives είναι σχετικά περιορισμένα σε σχέση με τον συνολικό αριθμό benign εγγραφών.

4.2.2 Πείραμα UNSW→UNSW

Στο δεύτερο πείραμα, τα μοντέλα εκπαιδεύτηκαν στο 70% του UNSW-NB15 και αξιολογήθηκαν στο υπόλοιπο 30%. Το UNSW-NB15 παρουσιάζει πιο ισορροπημένη κατανομή μεταξύ benign και attack κλάσεων σε σχέση με το CICIDS2017, γεγονός που επηρεάζει και τη συμπεριφορά των μοντέλων.

Και σε αυτό το πείραμα, το Random Forest είχε την καλύτερη απόδοση, με accuracy 93,2%, balanced accuracy 93,4% και F1-score 93,1%. Το precision ήταν 90,3% και το recall 96,0%, δείχνοντας ότι το μοντέλο διατήρησε ισχυρή απόδοση και στις δύο κλάσεις. Η Logistic Regression είχε επίσης καλή απόδοση, με F1-score 81,9% και recall 92,8%. Το LinearSVC πέτυχε F1-score 77,9%, με υψηλότερο precision αλλά χαμηλότερο recall. Σε σχέση με το CIC→CIC, τα γραμμικά μοντέλα αποδίδουν καλύτερα στο UNSW→UNSW, κυρίως επειδή η κατανομή των κλάσεων είναι πιο ισορροπημένη.



Σχήμα 3: Confusion Matrix του Random Forest στο UNSW→UNSW

Στο confusion matrix του Random Forest για το UNSW→UNSW παρατηρούνται τα εξής:

- True Negatives: 19.953
- False Positives: 2.052
- False Negatives: 790
- True Positives: 19.205

Το μοντέλο εντόπισε 19.205 από τις 19.995 επιθέσεις του test set, χάνοντας μόνο 790. Παρότι τα false positives είναι περισσότερα από το CIC→CIC, η συνολική εικόνα παραμένει ισχυρή, καθώς το μοντέλο επιτυγχάνει υψηλή ισορροπημένη απόδοση τόσο στην benign όσο και στην attack κλάση.

4.2.3 Συμπέρασμα Same-Dataset Πειραμάτων

Τα δύο same-dataset πειράματα δείχνουν ότι τα μοντέλα, και ειδικά το Random Forest, μπορούν να επιτύχουν πολύ υψηλή απόδοση όταν η εκπαίδευση και η αξιολόγηση γίνονται στο ίδιο dataset. Το Random Forest πέτυχε F1-score 95,2% στο CICIDS2017 και 93,1% στο UNSW-NB15.

Ωστόσο, τα αποτελέσματα αυτά δεν αρκούν για να αποδείξουν πραγματική επιχειρησιακή αξιοπιστία. Η αξιολόγηση εντός του ίδιου dataset ελέγχει κυρίως αν το μοντέλο μπορεί να μάθει μοτίβα που υπάρχουν στην ίδια κατανομή δεδομένων. Δεν ελέγχει αν το μοντέλο έχει μάθει γενικεύσιμα χαρακτηριστικά κακόβουλης συμπεριφοράς. Για τον λόγο αυτό, τα cross-dataset πειράματα είναι κρίσιμα.

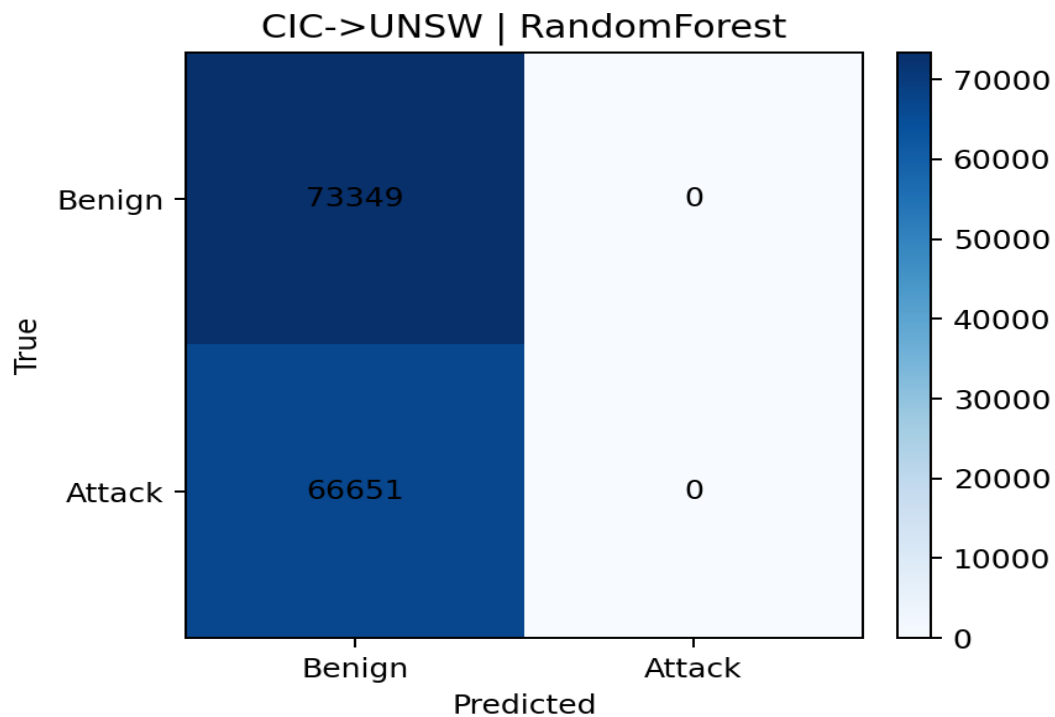
4.3 Cross-Dataset Αποτελέσματα

4.3.1 Πείραμα CIC→UNSW

Στο πείραμα CIC→UNSW, τα μοντέλα εκπαιδεύτηκαν σε ολόκληρο το CICIDS2017 και αξιολογήθηκαν σε ολόκληρο το UNSW-NB15. Το σενάριο αυτό εξετάζει αν η γνώση που αποκτάται από το CICIDS2017 μπορεί να μεταφερθεί στο UNSW-NB15.

Τα αποτελέσματα δείχνουν σχεδόν πλήρη αποτυχία γενίκευσης. Το Random Forest, ενώ είχε πετύχει F1-score 95,2% στο CIC→CIC, στο CIC→UNSW είχε F1-score 0,0%. Το μοντέλο δεν εντόπισε καμία επίθεση στο UNSW-NB15. Το balanced accuracy ήταν 50,0%, δηλαδή ισοδύναμο με τυχαία ή μονόπλευρη συμπεριφορά ταξινόμησης.

Η Logistic Regression είχε F1-score 3,1%, ενώ το LinearSVC είχε F1-score 1,5%. Παρότι τα δύο γραμμικά μοντέλα είχαν πολύ υψηλό precision, 91,4% και 95,3% αντίστοιχα, το recall τους ήταν εξαιρετικά χαμηλό. Αυτό σημαίνει ότι τα μοντέλα χαρακτήρισαν ελάχιστες εγγραφές ως attack. Όταν το έκαναν, συχνά είχαν δίκιο, αλλά απέτυχαν να εντοπίσουν τη συντριπτική πλειονότητα των πραγματικών επιθέσεων.



Σχήμα 4: Confusion Matrix του Random Forest στο CIC→UNSW

Στο confusion matrix του Random Forest για το CIC→UNSW παρατηρούνται τα εξής:

- True Negatives: 73.349
- False Positives: 0
- False Negatives: 66.651
- True Positives: 0

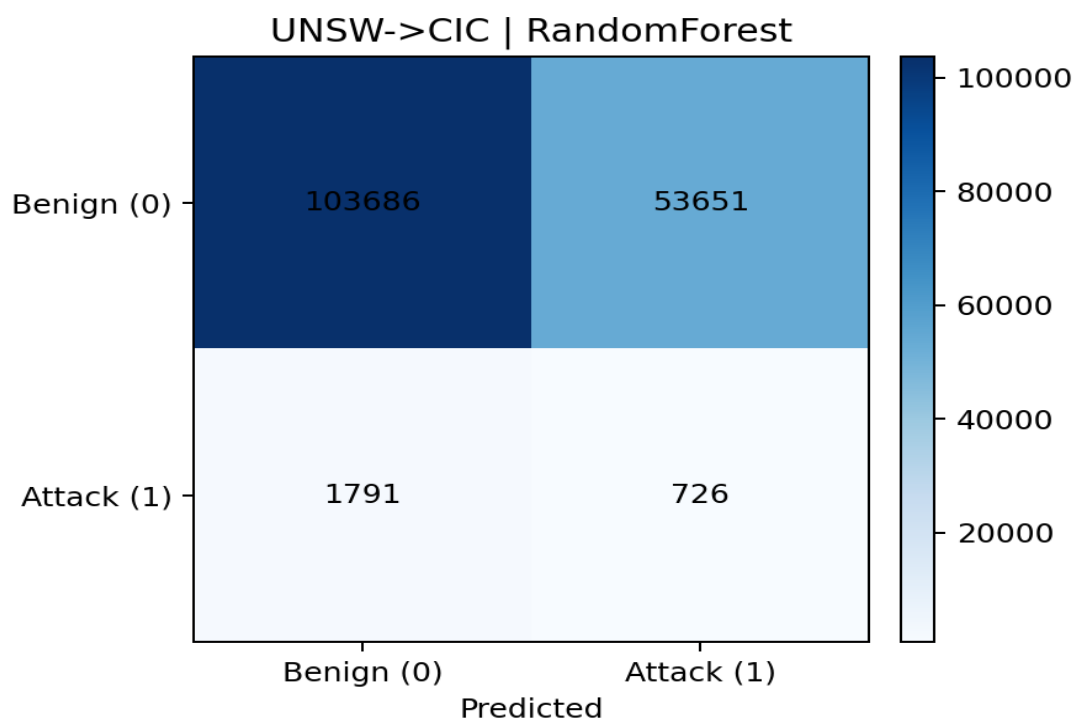
Το αποτέλεσμα αυτό είναι κρίσιμο. Το Random Forest ταξινόμησε όλες τις εγγραφές του UNSW-NB15 ως benign. Αυτό σημαίνει ότι δεν παρήγαγε false positives, αλλά ταυτόχρονα έχασε όλες τις επιθέσεις. Σε πραγματικό περιβάλλον IDS, ένα τέτοιο μοντέλο θα ήταν επικίνδυνο, επειδή θα δημιουργούσε ψευδή αίσθηση ασφάλειας ενώ δεν θα ανίχνευε καμία απειλή.

Η αποτυχία αυτή δείχνει ότι το Random Forest δεν έμαθε γενικεύσιμα μοτίβα επίθεσης. Αντίθετα, φαίνεται ότι έμαθε μοτίβα που είναι ειδικά για το CICIDS2017 και δεν ισχύουν στο UNSW-NB15.

4.3.2 Πείραμα UNSW→CIC

Στο πείραμα UNSW→CIC, τα μοντέλα εκπαιδεύτηκαν σε ολόκληρο το UNSW-NB15 και αξιολογήθηκαν σε ολόκληρο το CICIDS2017. Το σενάριο αυτό εξετάζει την αντίστροφη κατεύθυνση μεταφοράς γνώσης.

Και εδώ η απόδοση μειώθηκε δραματικά. Η Logistic Regression είχε F1-score 3,1%, το LinearSVC είχε F1-score 3,4%, ενώ το Random Forest είχε F1-score 17,2%. Το Random Forest είχε recall 66,2%, γεγονός που δείχνει ότι εντόπισε αρκετές από τις επιθέσεις του CICIDS2017. Ωστόσο, το precision ήταν μόλις 9,9%, επειδή παρήγαγε τεράστιο αριθμό false positives.



Σχήμα 5: Confusion Matrix του Random Forest στο UNSW→CIC

Στο confusion matrix του Random Forest για το UNSW→CIC παρατηρούνται τα εξής:

- True Negatives: 180.453
- False Positives: 325.369
- False Negatives: 18.164
- True Positives: 35.621

Το μοντέλο εντόπισε 35.621 επιθέσεις, αλλά ταυτόχρονα χαρακτήρισε λανθασμένα 325.369 benign εγγραφές ως attack. Ο λόγος false positives προς true positives είναι

περίπου 9,13. Αυτό σημαίνει ότι για κάθε πραγματική επίθεση που ανιχνεύεται, το μοντέλο παράγει περισσότερους από εννέα ψευδείς συναγερμούς.

Σε ένα πραγματικό Security Operations Center, αυτή η συμπεριφορά θα ήταν πρακτικά μη διαχειρίσιμη. Οι αναλυτές θα έπρεπε να εξετάζουν τεράστιο αριθμό λανθασμένων alarms, με αποτέλεσμα να μειώνεται η επιχειρησιακή χρησιμότητα του συστήματος.

4.3.3 Συμπέρασμα Cross-Dataset Πειραμάτων

Τα cross-dataset αποτελέσματα δείχνουν ότι όλα τα μοντέλα απέτυχαν να γενικεύσουν ικανοποιητικά. Το Random Forest, παρότι ήταν το καλύτερο μοντέλο στα same-dataset πειράματα, παρουσίασε τη μεγαλύτερη πτώση απόδοσης.

Στο CIC→UNSW απέτυχε σιωπηλά, ταξινομώντας όλες τις εγγραφές ως benign. Στο UNSW→CIC απέτυχε θορυβωδώς, παράγοντας τεράστιο αριθμό false positives. Αυτές οι δύο μορφές αποτυχίας είναι διαφορετικές αλλά εξίσου προβληματικές.

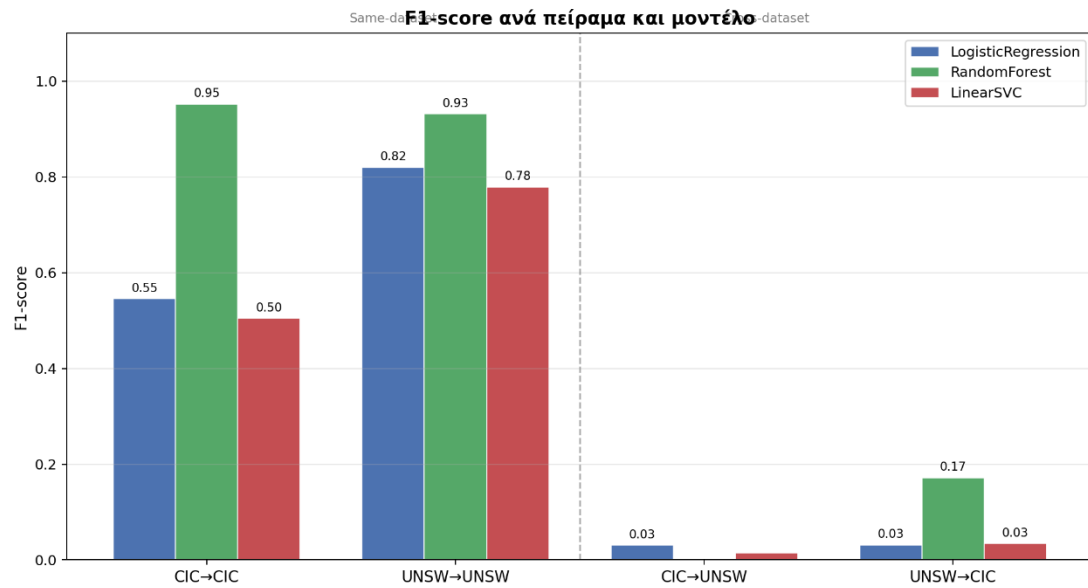
Το βασικό συμπέρασμα είναι ότι η υψηλή απόδοση εντός ενός dataset δεν μεταφέρεται αυτόματα σε διαφορετικό dataset. Τα μοντέλα δεν φαίνεται να μαθαίνουν καθολικά μοτίβα κακόβουλης συμπεριφοράς, αλλά χαρακτηριστικά που σχετίζονται με τη συγκεκριμένη κατανομή, τη μεθοδολογία συλλογής και τη διαδικασία feature extraction του dataset εκπαίδευσης.

4.4 Σύγκριση Same-Dataset και Cross-Dataset Απόδοσης

Η σύγκριση των same-dataset και cross-dataset αποτελεσμάτων αναδεικνύει το μέγεθος του generalization gap. Το πιο χαρακτηριστικό παράδειγμα είναι το Random Forest.

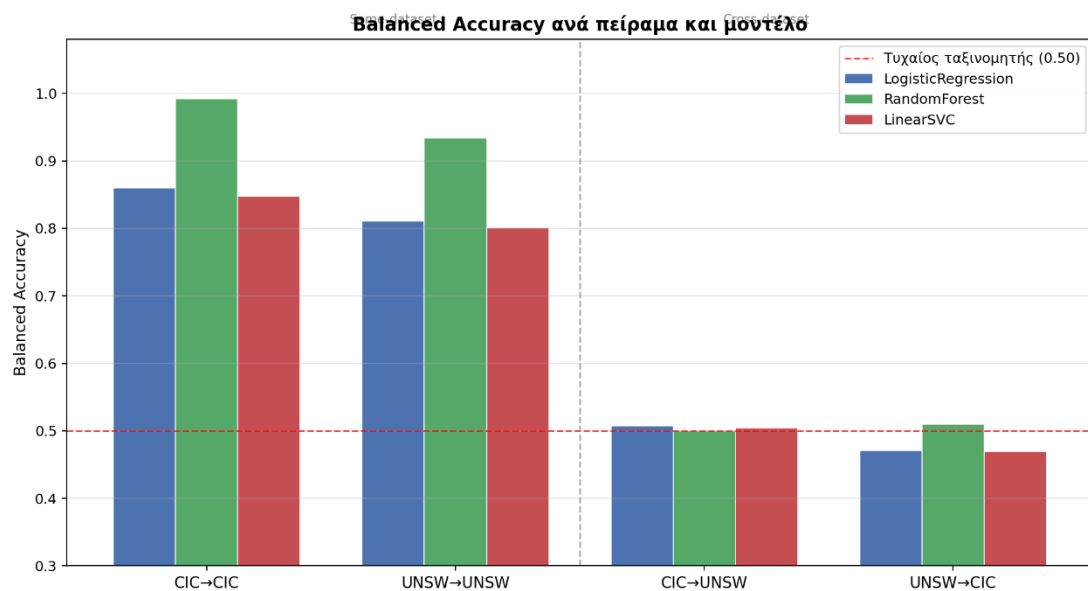
Στο CIC→CIC, το Random Forest είχε F1-score 95,2%. Στο CIC→UNSW, το F1-score έπεσε στο 0,0%. Η μείωση δF1 ήταν 0,952, δηλαδή πρακτικά πλήρης απώλεια της ανιχνευτικής ικανότητας.

Αντίστοιχα, στο UNSW→UNSW το Random Forest είχε F1-score 93,1%. Στο UNSW→CIC, το F1-score έπεσε στο 17,2%. Η μείωση δF1 ήταν 0,759. Παρότι το μοντέλο εντόπισε αρκετές επιθέσεις, η απόδοση παρέμεινε επιχειρησιακά αδύναμη λόγω του εξαιρετικά χαμηλού precision.



Σχήμα 6: Σύγκριση F1-score ανά πείραμα και μοντέλο

Η σύγκριση F1-score δείχνει καθαρά ότι τα same-dataset πειράματα βρίσκονται σε πολύ υψηλότερο επίπεδο απόδοσης σε σχέση με τα cross-dataset πειράματα. Ακόμη και το χειρότερο same-dataset αποτέλεσμα, δηλαδή το LinearSVC στο CIC→CIC με F1-score 50,4%, είναι πολύ υψηλότερο από τα περισσότερα cross-dataset αποτελέσματα.

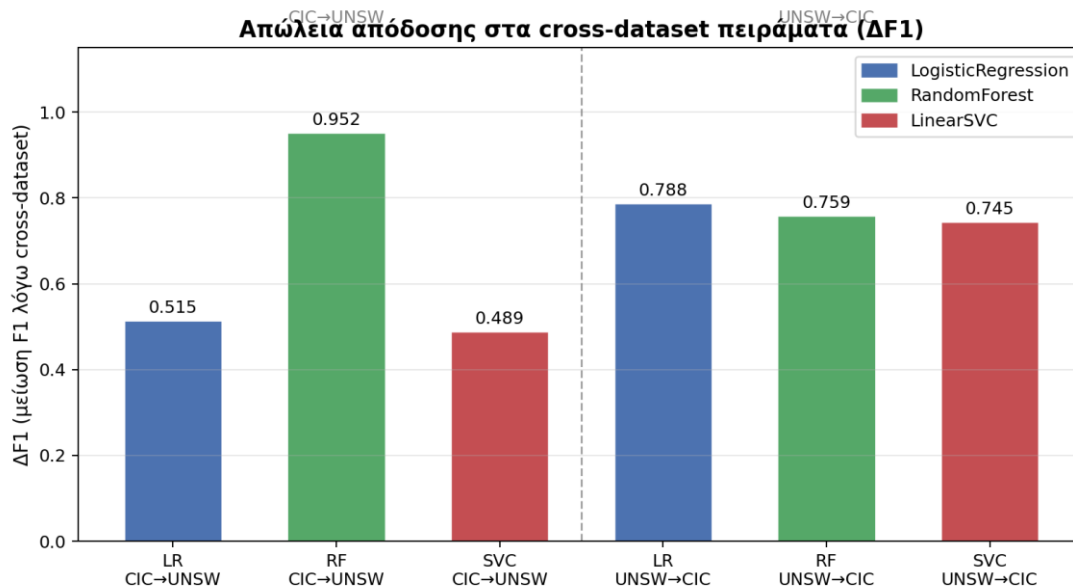


Σχήμα 7: Σύγκριση Balanced Accuracy ανά πείραμα και μοντέλο

Το balanced accuracy επιβεβαιώνει το ίδιο συμπέρασμα. Στα same-dataset πειράματα, το balanced accuracy κυμαίνεται από 80,1% έως 99,2%. Αντίθετα, στα cross-dataset

πειράματα οι τιμές βρίσκονται κοντά στο 50%, δηλαδή κοντά σε επίπεδο τυχαίας ταξινόμησης.

Η χρήση του balanced accuracy είναι ιδιαίτερα σημαντική, επειδή το απλό accuracy μπορεί να παραπλανήσει. Για παράδειγμα, στο UNSW→CIC η Logistic Regression έχει accuracy 82,9%, αλλά balanced accuracy μόλις 47,1% και F1-score 3,1%. Αυτό σημαίνει ότι το υψηλό accuracy οφείλεται κυρίως στην επικράτηση της benign κλάσης και όχι σε πραγματική ικανότητα ανίχνευσης επιθέσεων.



Σχήμα 8: Delta F1 στα cross-dataset σενάρια

Το Delta F1 αποτυπώνει άμεσα την πτώση απόδοσης. Οι μεγαλύτερες μειώσεις εμφανίζονται στο Random Forest, το οποίο ενώ είναι το ισχυρότερο μοντέλο στα same-dataset πειράματα, παρουσιάζει τη μεγαλύτερη κατάρρευση όταν αξιολογείται σε διαφορετικό dataset.

Αυτό το εύρημα είναι ιδιαίτερα σημαντικό. Δείχνει ότι τα πιο ισχυρά μοντέλα σε ένα benchmark δεν είναι απαραίτητα τα πιο αξιόπιστα σε πραγματικές συνθήκες μεταφοράς. Ένα μοντέλο μπορεί να εμφανίζεται εξαιρετικό σε ελεγχόμενο περιβάλλον, αλλά να αποτυγχάνει όταν αλλάξει η κατανομή των δεδομένων.

4.5 Ανάλυση Αποτυχίας ανά Μοντέλο

4.5.1 Logistic Regression

Η Logistic Regression είχε μέτρια έως καλή απόδοση στα same-dataset πειράματα, ειδικά στο UNSW→UNSW όπου πέτυχε F1-score 81,9%. Ωστόσο, στα cross-dataset πειράματα η απόδοσή της κατέρρευσε.

- Στο CIC→UNSW είχε F1-score 3,1%, με precision 91,4% αλλά recall μόλις 1,6%. Αυτό σημαίνει ότι το μοντέλο έκανε ελάχιστες θετικές προβλέψεις. Όταν χαρακτήριζε μία εγγραφή ως attack, συχνά ήταν σωστό, αλλά έχασε σχεδόν όλες τις πραγματικές επιθέσεις.
- Στο UNSW→CIC είχε επίσης F1-score 3,1%, με precision 3,4% και recall 2,9%. Η απόδοση αυτή δείχνει ότι το γραμμικό όριο απόφασης που έμαθε στο UNSW-NB15 δεν μπορούσε να διαχωρίσει αποτελεσματικά τις κλάσεις του CICIDS2017.

Η Logistic Regression επομένως απέτυχε κυρίως επειδή τα γραμμικά patterns που έμαθε στο source dataset δεν μεταφέρθηκαν στο target dataset.

4.5.2 Random Forest

Το Random Forest ήταν το καλύτερο μοντέλο στα same-dataset πειράματα, με F1-score 95,2% στο CIC→CIC και 93,1% στο UNSW→UNSW. Ωστόσο, στα cross-dataset πειράματα παρουσίασε την πιο χαρακτηριστική κατάρρευση.

- Στο CIC→UNSW, το F1-score ήταν 0,0%, καθώς το μοντέλο δεν εντόπισε καμία επίθεση. Αυτό υποδηλώνει ότι τα decision boundaries των δέντρων ήταν πλήρως προσαρμοσμένα στις κατανομές του CICIDS2017 και δεν ενεργοποιήθηκαν σωστά στο UNSW-NB15.
- Στο UNSW→CIC, το Random Forest είχε recall 66,2%, αλλά precision μόλις 9,9%. Αυτό σημαίνει ότι το μοντέλο ήταν υπερβολικά επιθετικό στις προβλέψεις attack, με αποτέλεσμα να παράγει τεράστιο αριθμό false positives.

Η συμπεριφορά αυτή δείχνει ότι το Random Forest μπορεί να αποδώσει εξαιρετικά όταν το train και test προέρχονται από την ίδια κατανομή, αλλά είναι ιδιαίτερα ευάλωτο σε dataset shift. Το μοντέλο φαίνεται να μαθαίνει πολύ συγκεκριμένα patterns του dataset εκπαίδευσης, τα οποία δεν παραμένουν αξιόπιστα όταν μεταφερθούν σε διαφορετικό περιβάλλον.

4.5.3 LinearSVC

Το LinearSVC είχε παρόμοια συμπεριφορά με τη Logistic Regression. Στα same-dataset πειράματα πέτυχε F1-score 50,4% στο CIC→CIC και 77,9% στο UNSW→UNSW. Η απόδοσή του ήταν χαμηλότερη από του Random Forest, αλλά παρέμεινε λειτουργική εντός του ίδιου dataset.

Στα cross-dataset πειράματα όμως απέτυχε σχεδόν πλήρως. Στο CIC→UNSW είχε F1-score 1,5%, ενώ στο UNSW→CIC είχε F1-score 3,4%. Το μοντέλο δεν κατάφερε να διατηρήσει ικανοποιητικό recall ούτε precision σε διαφορετική κατανομή δεδομένων.

Η αποτυχία του LinearSVC δείχνει ότι ακόμη και πιο απλά γραμμικά μοντέλα, τα οποία θεωρητικά μπορεί να είναι λιγότερο επιρρεπή σε overfitting σε σχέση με ensemble μοντέλα, δεν μπορούν να αντιμετωπίσουν από μόνα τους τη βαθιά ασυμβατότητα μεταξύ διαφορετικών IDS datasets.

4.6 Πρακτική Ερμηνεία των Αποτελεσμάτων

Τα αποτελέσματα έχουν σημαντική πρακτική σημασία για την ανάπτυξη ML-based IDS. Σε ένα πραγματικό περιβάλλον, ένα σύστημα ανίχνευσης εισβολών δεν αξιολογείται μόνο με βάση το αν πετυχαίνει υψηλή απόδοση σε ένα γνωστό benchmark. Πρέπει να μπορεί να λειτουργήσει αξιόπιστα σε διαφορετικά δίκτυα, διαφορετικές συσκευές, διαφορετικά traffic patterns και διαφορετικές κατανομές επιθέσεων.

Τα πειράματα δείχνουν ότι η αξιολόγηση μόνο εντός του ίδιου dataset μπορεί να είναι παραπλανητική. Το Random Forest φαίνεται εξαιρετικό στα same-dataset πειράματα, αλλά αποτυγχάνει σε cross-dataset συνθήκες. Στο CIC→UNSW δεν ανιχνεύει καμία επίθεση, ενώ στο UNSW→CIC παράγει υπερβολικά πολλούς ψευδείς συναγερμούς.

Αυτό δημιουργεί δύο διαφορετικούς επιχειρησιακούς κινδύνους:

1. **Σιωπηλή αποτυχία:** το IDS δεν ανιχνεύει επιθέσεις και τις χαρακτηρίζει ως benign. Αυτό παρατηρείται στο CIC→UNSW με το Random Forest.
2. **Θορυβώδης αποτυχία:** το IDS παράγει τεράστιο αριθμό false positives, καθιστώντας το μη διαχειρίσιμο από τους αναλυτές. Αυτό παρατηρείται στο UNSW→CIC με το Random Forest.

Και οι δύο μορφές αποτυχίας είναι σοβαρές. Η πρώτη οδηγεί σε απώλεια πραγματικών επιθέσεων. Η δεύτερη οδηγεί σε υπερφόρτωση του SOC και μείωση της εμπιστοσύνης στο σύστημα.

4.7 Σύνδεση με το Πρόβλημα του Dataset Shift

Η πτώση απόδοσης δεν μπορεί να εξηγηθεί μόνο από την επιλογή αλγορίθμου. Και τα τρία μοντέλα απέτυχαν στα cross-dataset σενάρια, παρότι ανήκουν σε διαφορετικές κατηγορίες ταξινομητών. Αυτό δείχνει ότι η αιτία είναι βαθύτερη και σχετίζεται με τη μετατόπιση κατανομής χαρακτηριστικών μεταξύ των datasets.

Το CICIDS2017 και το UNSW-NB15 διαφέρουν στη διαδικασία συλλογής, στα εργαλεία εξαγωγής flows, στις μονάδες μέτρησης, στην αναλογία benign/attack traffic και στη φύση των επιθέσεων. Ακόμη και όταν χρησιμοποιείται κοινό feature space, τα ίδια χαρακτηριστικά δεν έχουν απαραίτητα την ίδια στατιστική συμπεριφορά στα δύο datasets.

Για παράδειγμα, ένα χαρακτηριστικό όπως η διάρκεια ροής μπορεί να έχει διαφορετική μονάδα μέτρησης ή διαφορετική κατανομή. Αντίστοιχα, χαρακτηριστικά όπως τα bytes, τα packets και το rate επηρεάζονται από τον τρόπο παραγωγής και καταγραφής της κίνησης. Επομένως, ένα μοντέλο που εκπαιδεύεται σε ένα dataset μπορεί να μάθει decision boundaries που δεν έχουν νόημα στο άλλο dataset.

Η ανάλυση αυτή οδηγεί άμεσα στο επόμενο κεφάλαιο, όπου εξετάζονται πιο αναλυτικά οι αιτίες της αποτυχίας γενίκευσης μέσω feature distribution shift και SHAP explainability.

4.8 Σύνοψη Κεφαλαίου

Στο παρόν κεφάλαιο παρουσιάστηκαν τα πειραματικά αποτελέσματα της αξιολόγησης τριών μοντέλων Μηχανικής Μάθησης σε same-dataset και cross-dataset συνθήκες. Τα αποτελέσματα έδειξαν ότι τα μοντέλα μπορούν να επιτύχουν υψηλή απόδοση όταν εκπαιδεύονται και αξιολογούνται στο ίδιο dataset. Ειδικά το Random Forest πέτυχε F1-score 95,2% στο CIC→CIC και 93,1% στο UNSW→UNSW.

Ωστόσο, όταν τα ίδια μοντέλα αξιολογήθηκαν σε διαφορετικό dataset από αυτό της εκπαίδευσης, η απόδοση κατέρρευσε. Στο CIC→UNSW, το Random Forest δεν εντόπισε

καμία επίθεση, ενώ στο UNSW→CIC παρήγαγε τεράστιο αριθμό false positives. Τα γραμμικά μοντέλα παρουσίασαν επίσης πολύ χαμηλό F1-score στα cross-dataset σενάρια.

Τα ευρήματα επιβεβαιώνουν το βασικό ερευνητικό επιχείρημα της εργασίας: η υψηλή απόδοση ενός ML-based IDS σε ένα μεμονωμένο dataset δεν αποτελεί επαρκή ένδειξη πραγματικής ικανότητας γενίκευσης. Τα μοντέλα φαίνεται να μαθαίνουν σε μεγάλο βαθμό τα ιδιαίτερα χαρακτηριστικά του dataset εκπαίδευσης, αντί για καθολικά μοτίβα κακόβουλης δικτυακής συμπεριφοράς.

Το επόμενο κεφάλαιο εστιάζει στην ερμηνεία αυτής της αποτυχίας, εξετάζοντας τη μετατόπιση κατανομής των χαρακτηριστικών και την ανάλυση SHAP για το Random Forest.

5.Ανάλυση Αιτιών και Ερμηνεία Αποτελεσμάτων

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν τα πειραματικά αποτελέσματα της αξιολόγησης των μοντέλων Logistic Regression, Random Forest και LinearSVC σε same-dataset και cross-dataset σενάρια. Τα αποτελέσματα έδειξαν έντονη αντίθεση μεταξύ της απόδοσης εντός του ίδιου dataset και της απόδοσης σε διαφορετικό dataset. Συγκεκριμένα, το Random Forest πέτυχε πολύ υψηλή απόδοση στα same-dataset πειράματα, με F1-score 0,952 στο CIC→CIC και 0,931 στο UNSW→UNSW, αλλά κατέρρευσε στα cross-dataset σενάρια, με F1-score 0,000 στο CIC→UNSW και 0,172 στο UNSW→CIC.

Το παρόν κεφάλαιο επιχειρεί να ερμηνεύσει τους λόγους αυτής της κατάρρευσης. Η ανάλυση δεν περιορίζεται στη διαπίστωση ότι τα μοντέλα απέτυχαν, αλλά εξετάζει γιατί απέτυχαν. Για τον σκοπό αυτό αξιοποιούνται δύο βασικές προσεγγίσεις:

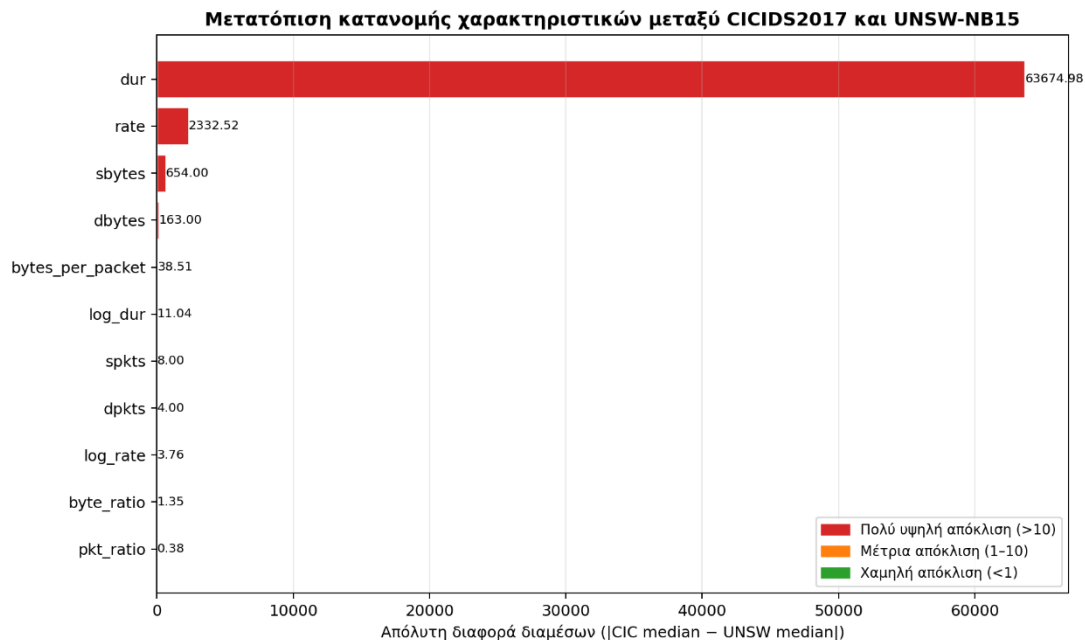
- Πρώτον, πραγματοποιείται ανάλυση μετατόπισης κατανομής χαρακτηριστικών μεταξύ CICIDS2017 και UNSW-NB15. Η ανάλυση αυτή εξετάζει κατά πόσο τα ίδια χαρακτηριστικά έχουν παρόμοια ή διαφορετική στατιστική συμπεριφορά στα δύο datasets.
- Δεύτερον, εφαρμόζεται ανάλυση Explainable AI μέσω SHAP στο Random Forest. Η SHAP ανάλυση χρησιμοποιείται ως εργαλείο ερμηνείας της συμπεριφοράς του μοντέλου, αλλά κυρίως ως μηχανισμός ανάλυσης αποτυχίας. Ο στόχος δεν είναι να αποδειχθεί ότι

το μοντέλο αποφασίζει σωστά, αλλά να εντοπιστεί ποια χαρακτηριστικά επηρεάζουν τις αποφάσεις του και κατά πόσο αυτά παραμένουν σταθερά μεταξύ διαφορετικών datasets.

5.2 Μετατόπιση Κατανομής Χαρακτηριστικών

Η βασική υπόθεση πίσω από τα cross-dataset πειράματα είναι ότι ένα μοντέλο που έχει μάθει γενικεύσιμα μοτίβα κακόβουλης συμπεριφοράς θα πρέπει να διατηρεί ένα ελάχιστο αποδεκτό επίπεδο απόδοσης όταν εφαρμόζεται σε διαφορετικό dataset. Ωστόσο, τα αποτελέσματα του Κεφαλαίου 5 έδειξαν ότι αυτό δεν συνέβη. Η πρώτη πιθανή εξήγηση είναι ότι τα δύο datasets δεν μοιράζονται πραγματικά το ίδιο feature distribution, ακόμη και όταν χρησιμοποιούνται χαρακτηριστικά με παρόμοια σημασιολογική έννοια.

Για τον λόγο αυτό πραγματοποιήθηκε στατιστική σύγκριση των 11 χαρακτηριστικών που χρησιμοποιήθηκαν στην εργασία. Για κάθε χαρακτηριστικό υπολογίστηκαν η μέση τιμή, η διάμεσος, το ενδοτεταρτημοριακό εύρος και οι απόλυτες διαφορές μεταξύ CICIDS2017 και UNSW-NB15. Η διαδικασία αυτή επιτρέπει να εξεταστεί αν η αποτυχία γενίκευσης οφείλεται μόνο στους αλγορίθμους ή αν προκύπτει από βαθύτερη ασυμβατότητα των δεδομένων.



Σχήμα 9: Feature Distribution Shift μεταξύ CICIDS2017 και UNSW-NB15

Πίνακας 5: Στατιστική σύγκριση χαρακτηριστικών μεταξύ CICIDS2017 και UNSW-NB15

Χαρακτηρι- στικό	CIC Median	UNSW Median	Διαφορά Διαμέσων	CIC IQR	UNSW IQR
dur	63.675,000	0,023	63.674,977	5.650.807,500	0,676
rate	54,708	2.387,230	2.332,522	11.797,760	99.963,684
sbytes	64,000	718,000	654,000	326,000	2.006,000
dbytes	141,000	304,000	163,000	796,000	2.478,000
bytes_per_p acket	24,000	62,513	38,513	32,417	42,043
log_dur	11,062	0,023	11,039	9,987	0,516
spkts	2,000	10,000	8,000	4,000	18,000
dpkts	2,000	6,000	4,000	4,000	18,000
log_rate	4,020	7,778	3,758	8,400	7,893
byte_ratio	0,459	1,814	1,355	0,824	113,216
pkt_ratio	0,736	1,111	0,375	0,837	1,120

Τα αποτελέσματα δείχνουν ότι υπάρχουν έντονες διαφορές σε κρίσιμα χαρακτηριστικά. Το πιο χαρακτηριστικό παράδειγμα είναι το dur. Στο CICIDS2017 η διάμεσος της διάρκειας ροής είναι 63.675, ενώ στο UNSW-NB15 είναι 0,023. Η διαφορά αυτή δεν είναι απλώς στατιστική· συνδέεται με το γεγονός ότι η διάρκεια ροής στα δύο datasets εκφράζεται σε διαφορετικές μονάδες μέτρησης. Στο CICIDS2017 το Flow Duration καταγράφεται σε μικροδευτερόλεπτα, ενώ στο UNSW-NB15 το dur καταγράφεται σε δευτερόλεπτα. Η διαφορά αυτή έχει ήδη αναγνωριστεί ως βασικό εμπόδιο στη σύγκριση των δύο datasets.

Σημαντική απόκλιση εμφανίζεται και στο rate. Η διάμεσος του rate στο CICIDS2017 είναι 54,708, ενώ στο UNSW-NB15 είναι 2.387,230. Αυτό σημαίνει ότι το UNSW-NB15 εμφανίζει πολύ υψηλότερους ρυθμούς πακέτων ανά δευτερόλεπτο. Η διαφορά αυτή μπορεί να οφείλεται τόσο σε διαφορετικές συνθήκες παραγωγής δικτυακής κίνησης όσο και σε διαφορετική μεθοδολογία εξαγωγής χαρακτηριστικών.

Αντίστοιχα, τα sbytes, dbytes, spkts και dpkts παρουσιάζουν αισθητές διαφορές. Το UNSW-NB15 έχει υψηλότερες διαμέσους σε πακέτα και bytes, γεγονός που δείχνει ότι οι ροές του έχουν διαφορετική δομή από τις αντίστοιχες ροές του CICIDS2017. Ακόμη

και τα engineered features, όπως το bytes_per_packet, το byte_ratio και το pkt_ratio, δεν παραμένουν σταθερά μεταξύ των δύο datasets.

Η σημαντικότερη παρατήρηση είναι ότι η μετατόπιση δεν αφορά μόνο ένα χαρακτηριστικό. Αν το πρόβλημα περιοριζόταν αποκλειστικά στο dur, τότε θα μπορούσε να θεωρηθεί απλό ζήτημα μονάδων μέτρησης. Ωστόσο, η απόκλιση εμφανίζεται σε πολλά χαρακτηριστικά ταυτόχρονα. Αυτό δείχνει ότι τα δύο datasets έχουν διαφορετικό στατιστικό προφίλ και, επομένως, διαφορετικό feature distribution.

5.3 Ο Ρόλος του RobustScaler

Στη μεθοδολογία χρησιμοποιήθηκε RobustScaler, καθώς είναι πιο ανθεκτικός σε outliers σε σχέση με μεθόδους που βασίζονται στη μέση τιμή και την τυπική απόκλιση. Ωστόσο, τα αποτελέσματα δείχνουν ότι η χρήση RobustScaler δεν επαρκεί για να λύσει το πρόβλημα της cross-dataset μεταφοράς.

Ο RobustScaler εκτελεί τον μετασχηματισμό:

$$x_scaled = (x - median_train) / IQR_train$$

Αυτό σημαίνει ότι ο scaler μαθαίνει τη διάμεσο και το IQR από το dataset εκπαίδευσης και στη συνέχεια εφαρμόζει τον ίδιο μετασχηματισμό στο dataset αξιολόγησης. Στα same-dataset πειράματα αυτή η διαδικασία είναι λογική, επειδή train και test προέρχονται από την ίδια κατανομή. Στα cross-dataset πειράματα όμως, το πρόβλημα γίνεται πολύ πιο έντονο.

Στο σενάριο CIC→UNSW, ο scaler εκπαιδεύεται στα στατιστικά του CICIDS2017 και εφαρμόζεται στο UNSW-NB15. Για το χαρακτηριστικό dur, η διάμεσος του CICIDS2017 είναι 63.675 και το IQR είναι 5.650.807,5, ενώ μία τυπική τιμή του UNSW-NB15 είναι 0,023. Επομένως, μετά την εφαρμογή του RobustScaler, οι τιμές του UNSW-NB15 τοποθετούνται σε χώρο που δεν αντιστοιχεί στη σημασία που είχαν οι τιμές του CICIDS2017 κατά την εκπαίδευση. Στην πράξη, το μοντέλο βλέπει τη διάρκεια των ροών του UNSW-NB15 μέσα από τη στατιστική κλίμακα του CICIDS2017. Το υπάρχον draft περιγράφει ακριβώς αυτό το πρόβλημα για το dur, δείχνοντας ότι η scaled τιμή γίνεται πρακτικά πολύ κοντά στο μηδέν και άρα χάνει την πραγματική σημασία της για το μοντέλο.

Αυτό εξηγεί γιατί ένα μοντέλο όπως το Random Forest μπορεί να αποτύχει πλήρως στο CIC→UNSW. Τα decision boundaries των δέντρων έχουν δημιουργηθεί με βάση τις κλιμακωμένες τιμές του CICIDS2017. Όταν οι τιμές του UNSW-NB15 μετασχηματίζονται με βάση τα στατιστικά του CICIDS2017, δεν τοποθετούνται απαραίτητα σε περιοχές του feature space όπου το μοντέλο έχει μάθει να αναγνωρίζει επιθέσεις.

Το ίδιο φαινόμενο λειτουργεί και αντίστροφα στο UNSW→CIC. Εκεί ο scaler μαθαίνει τα στατιστικά του UNSW-NB15 και εφαρμόζεται στο CICIDS2017. Το αποτέλεσμα είναι διαφορετικό είδος αποτυχίας: το Random Forest δεν ταξινομεί τα πάντα ως benign, αλλά παράγει τεράστιο αριθμό false positives. Αυτό δείχνει ότι η κατεύθυνση μεταφοράς επηρεάζει τον τρόπο αποτυχίας του μοντέλου.

Η χρήση RobustScaler επομένως είναι μεθοδολογικά σωστή για την αποφυγή διαρροής δεδομένων, αλλά δεν αρκεί για να εξαλείψει τη μετατόπιση κατανομής. Το πρόβλημα δεν είναι απλώς η ύπαρξη outliers. Είναι ότι τα δύο datasets έχουν διαφορετικές στατιστικές δομές.

5.4 Ερμηνεία της Αποτυχίας CIC→UNSW

Στο πείραμα CIC→UNSW, το Random Forest εκπαιδεύτηκε σε ολόκληρο το CICIDS2017 και αξιολογήθηκε στο UNSW-NB15. Το αποτέλεσμα ήταν πλήρης αποτυχία ανίχνευσης επιθέσεων, καθώς το μοντέλο ταξινόμησε όλες τις εγγραφές του UNSW-NB15 ως benign. Το confusion matrix έδειξε 73.349 true negatives, 0 false positives, 66.651 false negatives και 0 true positives.

Αυτός ο τύπος αποτυχίας μπορεί να χαρακτηριστεί ως **σιωπηλή αποτυχία**. Το μοντέλο δεν δημιουργεί ψευδείς συναγερμούς, αλλά αυτό δεν είναι ένδειξη καλής λειτουργίας. Αντίθετα, είναι ένδειξη ότι το μοντέλο δεν αναγνωρίζει καμία κακόβουλη δραστηριότητα στο target dataset.

Η εξήγηση συνδέεται με το γεγονός ότι το Random Forest έχει μάθει συγκεκριμένα μοτίβα του CICIDS2017. Στο CICIDS2017, τα byte-related χαρακτηριστικά και η διάρκεια ροής φαίνεται να έχουν συγκεκριμένη σχέση με την attack κλάση. Όταν όμως το μοντέλο εφαρμόζεται στο UNSW-NB15, οι τιμές αυτών των χαρακτηριστικών δεν αντιστοιχούν στα ίδια μοτίβα.

Η αποτυχία αυτή δεν σημαίνει ότι το Random Forest είναι γενικά κακός αλγόριθμος. Αντίθετα, στο CIC→CIC είχε εξαιρετική απόδοση. Αυτό που δείχνει είναι ότι η υψηλή same-dataset απόδοση μπορεί να είναι αποτέλεσμα προσαρμογής σε dataset-specific patterns. Όταν αυτά τα patterns δεν υπάρχουν ή δεν έχουν την ίδια στατιστική σημασία στο target dataset, το μοντέλο αποτυγχάνει.

5.5 Ερμηνεία της Αποτυχίας UNSW→CIC

Στο πείραμα UNSW→CIC, το Random Forest εκπαιδεύτηκε στο UNSW-NB15 και αξιολογήθηκε στο CICIDS2017. Σε αντίθεση με το CIC→UNSW, εδώ το μοντέλο δεν ταξινομήσε όλες τις εγγραφές ως benign. Αντίθετα, εντόπισε αρκετές επιθέσεις, αλλά παρήγαγε τεράστιο αριθμό false positives.

Το confusion matrix έδειξε 180.453 true negatives, 325.369 false positives, 18.164 false negatives και 35.621 true positives. Το recall ήταν 66,2%, αλλά το precision μόλις 9,9%. Αυτό σημαίνει ότι το μοντέλο είχε την τάση να χαρακτηρίζει υπερβολικά πολλές εγγραφές ως επιθέσεις.

Αυτός ο τύπος αποτυχίας μπορεί να χαρακτηριστεί ως **θορυβώδης αποτυχία**. Το μοντέλο ανιχνεύει ένα μέρος των πραγματικών επιθέσεων, αλλά δημιουργεί τόσο πολλούς ψευδείς συναγερμούς ώστε γίνεται πρακτικά μη χρήσιμο. Σε πραγματικό περιβάλλον SOC, ένας τόσο μεγάλος αριθμός false positives θα οδηγούσε σε alarm fatigue και σε απώλεια εμπιστοσύνης στο σύστημα.

Η διαφορά ανάμεσα στο CIC→UNSW και στο UNSW→CIC είναι σημαντική. Στο πρώτο σενάριο, το μοντέλο είναι υπερβολικά συντηρητικό και δεν ανιχνεύει επιθέσεις. Στο δεύτερο σενάριο, το μοντέλο είναι υπερβολικά επιθετικό και χαρακτηρίζει μεγάλο μέρος της benign κίνησης ως κακόβουλη. Αυτό δείχνει ότι η αποτυχία γενίκευσης δεν έχει πάντα την ίδια μορφή. Εξαρτάται από το source dataset, το target dataset και τον τρόπο με τον οποίο οι κατανομές τους επικαλύπτονται ή αποκλίνουν.

5.6 SHAP-Based Explainable Failure Analysis

Η ανάλυση SHAP εφαρμόστηκε στο Random Forest, καθώς ήταν το μοντέλο με την υψηλότερη απόδοση στα same- dataset πειράματα και ταυτόχρονα παρουσίασε την πιο χαρακτηριστική κατάρρευση στα cross- dataset σενάρια. Η επιλογή του SHAP είναι κατάλληλη, επειδή επιτρέπει την ποσοτική αποτίμηση της συνεισφοράς κάθε

χαρακτηριστικού στις αποφάσεις του μοντέλου. Στη βιβλιογραφική ανασκόπηση είχε ήδη αναδειχθεί ότι το SHAP μπορεί να χρησιμοποιηθεί για την αποκάλυψη patterns που δεν είναι εύκολα ορατά με παραδοσιακές μετρικές.

Στην παρούσα εργασία, το SHAP δεν χρησιμοποιείται ως εργαλείο απόδειξης ορθότητας του μοντέλου. Αντίθετα, χρησιμοποιείται ως εργαλείο **explainable failure analysis**. Αυτό είναι κρίσιμο, επειδή στα cross-dataset σενάρια τα μοντέλα δεν έχουν καλή απόδοση. Επομένως, το SHAP δεν απαντά στο ερώτημα “γιατί το μοντέλο αποφάσισε σωστά;”, αλλά στο ερώτημα “ποια χαρακτηριστικά επηρεάζουν τις αποφάσεις ενός μοντέλου που αποτυγχάνει να γενικεύσει;”.

Η ανάλυση πραγματοποιήθηκε με βάση τη μέση απόλυτη SHAP τιμή κάθε χαρακτηριστικού. Όσο μεγαλύτερη είναι η μέση απόλυτη SHAP τιμή, τόσο μεγαλύτερη είναι η συνολική επίδραση του χαρακτηριστικού στις προβλέψεις του μοντέλου.

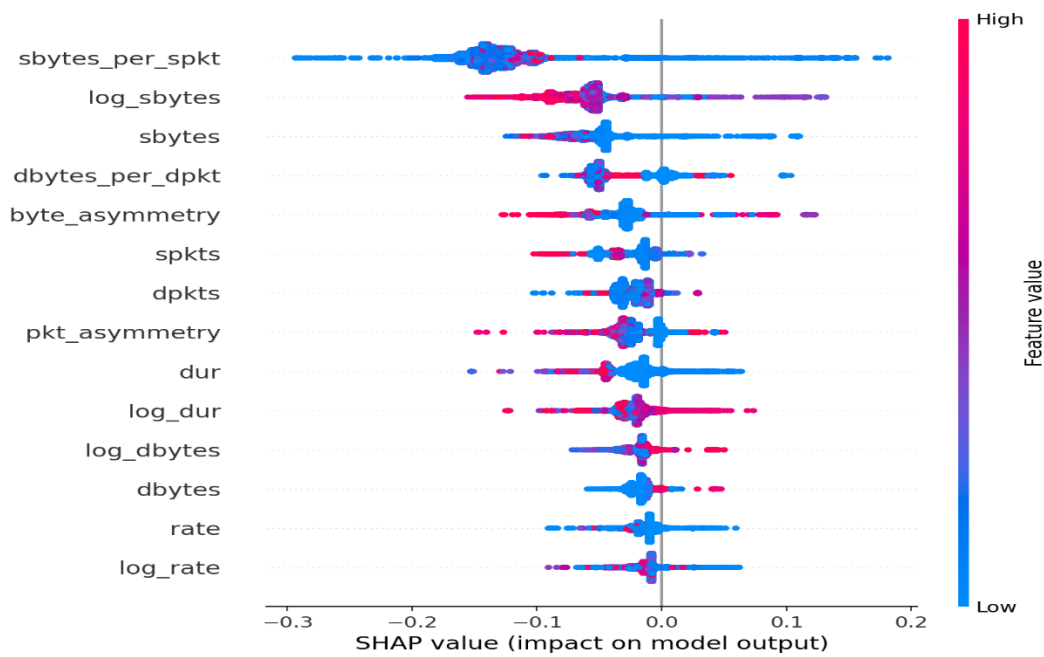
5.7 SHAP Ανάλυση στα Same- Dataset Πειράματα

5.7.1 CIC→CIC

Στο πείραμα CIC→CIC, τα σημαντικότερα χαρακτηριστικά για το Random Forest ήταν:

Κατάταξη	Χαρακτηριστικό	Mean Absolute SHAP
1	bytes_per_packet	0,1456
2	sbytes	0,0773
3	byte_ratio	0,0711
4	dbytes	0,0490
5	spkts	0,0392

Η κατάταξη αυτή δείχνει ότι στο CICIDS2017 το Random Forest βασίζεται κυρίως σε χαρακτηριστικά που σχετίζονται με τον όγκο δεδομένων και τη δομή των bytes ανά flow. Το bytes_per_packet, το sbytes και το byte_ratio κυριαρχούν στις εξηγήσεις. Αυτό σημαίνει ότι το μοντέλο αναγνωρίζει επιθέσεις με βάση μοτίβα που σχετίζονται με το μέγεθος των πακέτων, την ασυμμετρία των bytes και τη συνολική ποσότητα δεδομένων προς την κατεύθυνση της πηγής.



Σχήμα 10: SHAP Summary Plot για το Random Forest στο CIC→CIC

Η υψηλή απόδοση του Random Forest στο CIC→CIC φαίνεται να οφείλεται στην ικανότητά του να αξιοποιεί μη γραμμικές σχέσεις μεταξύ byte-related χαρακτηριστικών. Ωστόσο, αυτό δημιουργεί και τον βασικό κίνδυνο: αν τα συγκεκριμένα byte patterns είναι χαρακτηριστικά του CICIDS2017 και όχι γενικεύσιμα μοτίβα επιθέσεων, τότε το μοντέλο θα δυσκολευτεί να μεταφερθεί σε άλλο dataset.

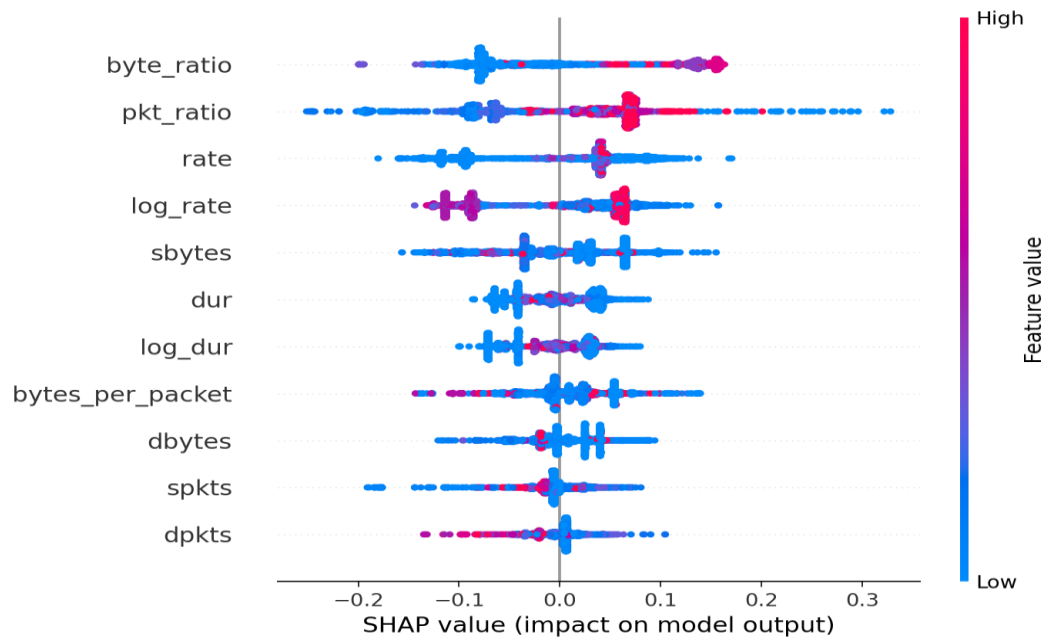
5.7.2 UNSW→UNSW

Στο πείραμα UNSW→UNSW, τα σημαντικότερα χαρακτηριστικά ήταν:

Κατάταξη	Χαρακτηριστικό	Mean Absolute SHAP
1	byte_ratio	0,0831
2	pkt_ratio	0,0704
3	rate	0,0674
4	log_rate	0,0652
5	sbytes	0,0428

Στο UNSW-NB15, η εικόνα είναι διαφορετική. Το μοντέλο βασίζεται περισσότερο σε αναλογικά χαρακτηριστικά και σε χαρακτηριστικά ρυθμού, όπως byte_ratio, pkt_ratio,

rate και log_rate. Αυτό δείχνει ότι η διάκριση benign/attack στο UNSW-NB15 δεν στηρίζεται στα ίδια χαρακτηριστικά με το CICIDS2017.



Σχήμα 11: SHAP Summary Plot για το Random Forest στο UNSW→UNSW

Η σύγκριση CIC→CIC και UNSW→UNSW είναι αποκαλυπτική. Στο CICIDS2017 κυριαρχεί το bytes_per_packet, ενώ στο UNSW-NB15 κυριαρχούν τα byte_ratio, pkt_ratio και rate. Επομένως, ακόμη και όταν το Random Forest έχει υψηλή απόδοση και στα δύο datasets, δεν χρησιμοποιεί τα ίδια χαρακτηριστικά με τον ίδιο τρόπο. Αυτό αποτελεί ένδειξη ότι τα δύο datasets έχουν διαφορετικούς μηχανισμούς διαχωρισμού μεταξύ benign και attack traffic.

5.8 SHAP Ανάλυση στα Cross-Dataset Πειράματα

5.8.1 CIC→UNSW

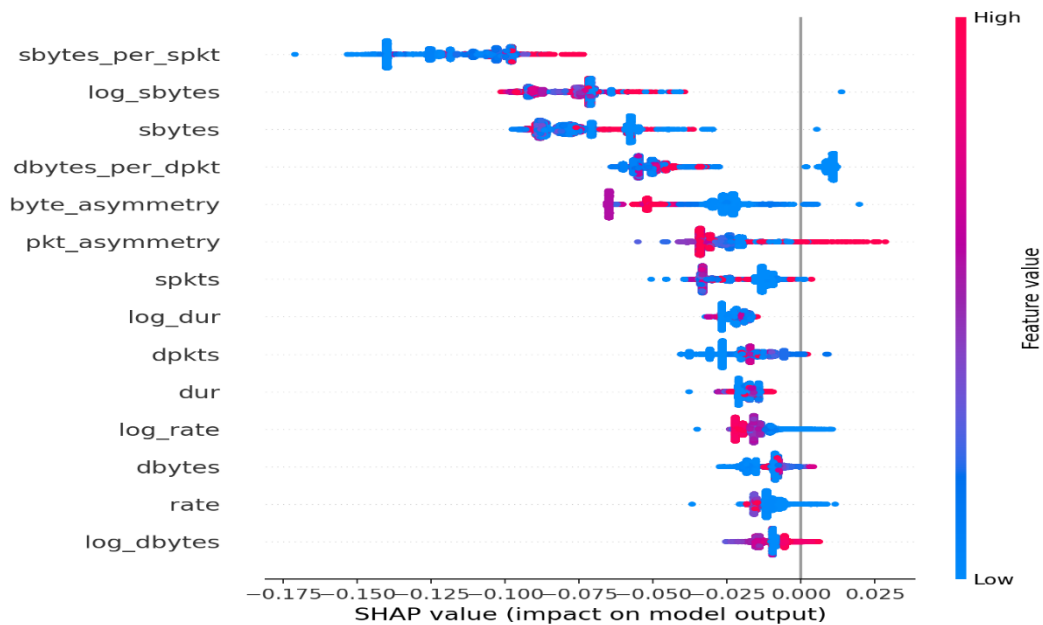
Στο πείραμα CIC→UNSW, τα σημαντικότερα χαρακτηριστικά σύμφωνα με το SHAP ήταν:

Κατάταξη	Χαρακτηριστικό	Mean Absolute SHAP
1	bytes_per_packet	0,1374
2	sbytes	0,1004

Κατάταξη	Χαρακτηριστικό	Mean Absolute SHAP
3	byte_ratio	0,0600
4	spkts	0,0392
5	log_dur	0,0321

Η κατάταξη αυτή είναι πολύ κοντά στη λογική του CIC→CIC. Το μοντέλο συνεχίζει να δίνει μεγάλη σημασία στα bytes_per_packet, sbytes και byte_ratio. Όμως, στο CIC→UNSW το Random Forest είχε F1-score 0,000 και δεν εντόπισε καμία επίθεση.

Αυτό είναι κρίσιμο εύρημα. Το SHAP δείχνει ότι το μοντέλο εξακολουθεί να εφαρμόζει τη λογική που έμαθε από το CICIDS2017, αλλά αυτή η λογική δεν λειτουργεί στο UNSW-NB15. Δηλαδή, τα χαρακτηριστικά που ήταν χρήσιμα για το CICIDS2017 δεν μεταφέρονται ως αξιόπιστοι δείκτες επίθεσης στο UNSW-NB15.



Σχήμα 12: SHAP Summary Plot για το Random Forest στο CIC→UNSW

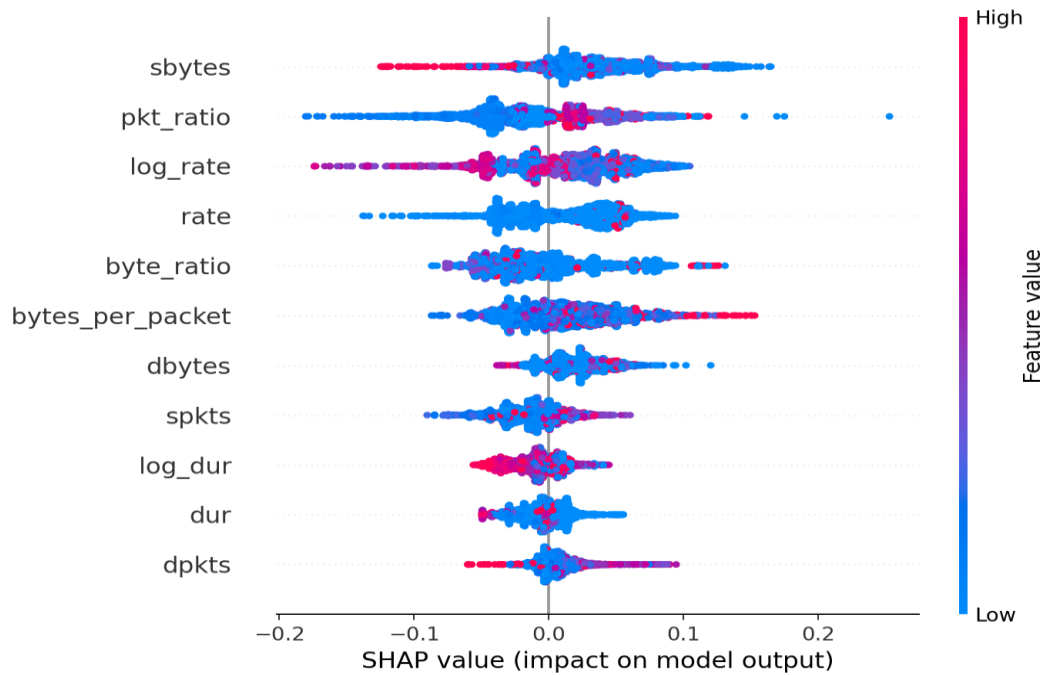
Επομένως, η SHAP ανάλυση δεν δείχνει επιτυχημένη ερμηνεία σωστών αποφάσεων. Δείχνει αποτυχία μεταφοράς. Το μοντέλο βασίζεται σε χαρακτηριστικά που γνωρίζει από το source dataset, αλλά το target dataset δεν ακολουθεί την ίδια στατιστική σχέση μεταξύ αυτών των χαρακτηριστικών και της attack κλάσης.

5.8.2 UNSW→CIC

Στο πείραμα UNSW→CIC, τα σημαντικότερα χαρακτηριστικά ήταν:

Κατάταξη	Χαρακτηριστικό	Mean Absolute SHAP
1	sbytes	0,0389
2	pkt_ratio	0,0380
3	log_rate	0,0356
4	rate	0,0340
5	byte_ratio	0,0308

Σε αυτό το σενάριο, το μοντέλο βασίζεται σε συνδυασμό χαρακτηριστικών bytes, packet ratios και rate. Ωστόσο, όπως φάνηκε στο Κεφάλαιο 5, η απόδοση παραμένει χαμηλή, επειδή το precision είναι μόλις 9,9%. Το μοντέλο παράγει τεράστιο αριθμό false positives.



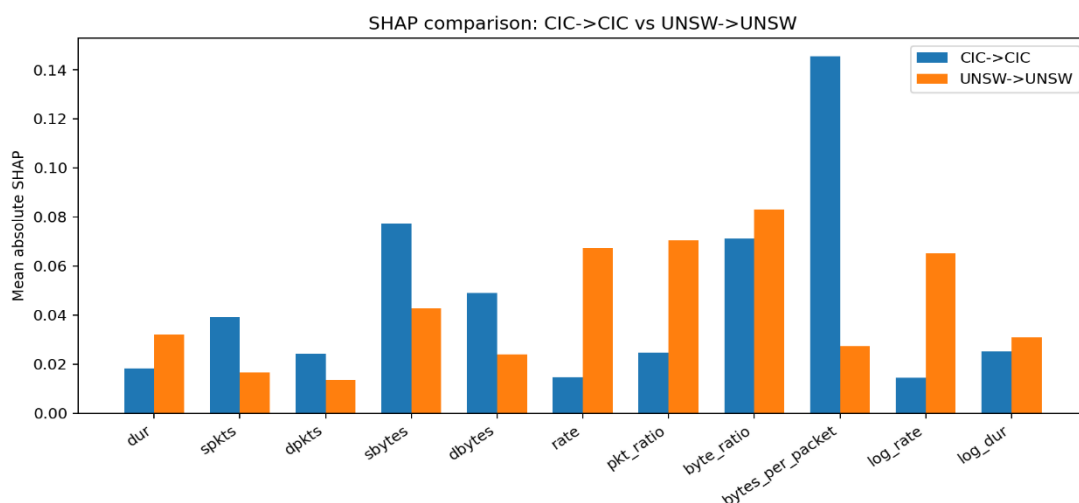
Σχήμα 13: SHAP Summary Plot για το Random Forest στο UNSW→CIC

Η SHAP ανάλυση δείχνει ότι το μοντέλο προσπαθεί να εφαρμόσει patterns που έμαθε από το UNSW-NB15 στο CICIDS2017. Όμως, στο CICIDS2017 τα ίδια χαρακτηριστικά δεν διαχωρίζουν με τον ίδιο τρόπο την benign από την attack κίνηση. Το αποτέλεσμα είναι υπερβολικά επιθετική ταξινόμηση και μαζική παραγωγή ψευδών συναγερμών.

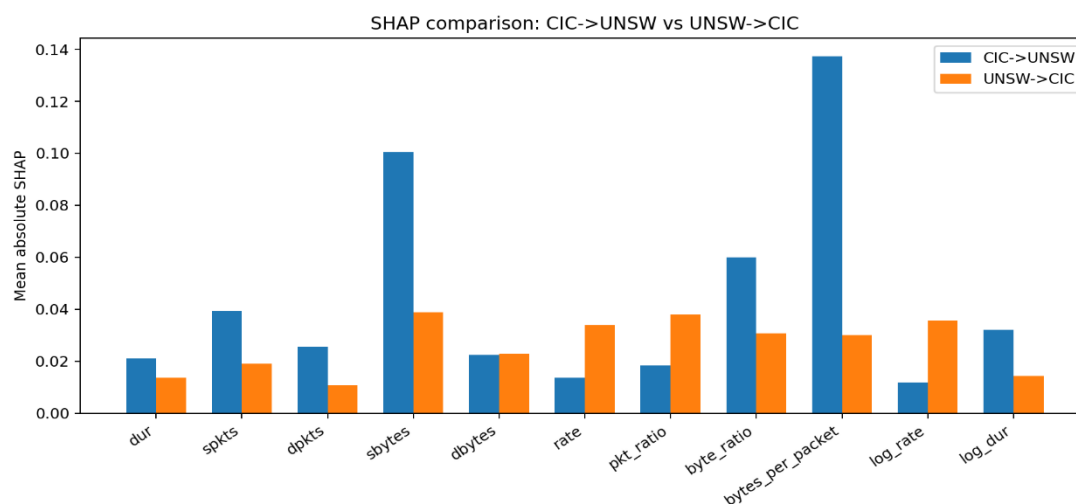
5.9 Σύγκριση SHAP Σημαντικότητας Μεταξύ Πειραμάτων

Η σύγκριση των SHAP αποτελεσμάτων μεταξύ same-dataset και cross-dataset πειραμάτων οδηγεί σε τρία βασικά συμπεράσματα.

- Πρώτον, τα σημαντικότερα χαρακτηριστικά δεν είναι σταθερά μεταξύ CICIDS2017 και UNSW-NB15. Στο CIC→CIC, το κορυφαίο χαρακτηριστικό είναι το bytes_per_packet, ενώ στο UNSW→UNSW είναι το byte_ratio. Αυτό δείχνει ότι κάθε dataset έχει διαφορετική εσωτερική δομή και διαφορετικά χαρακτηριστικά που οδηγούν την ταξινόμηση.
- Δεύτερον, στα cross-dataset πειράματα το μοντέλο διατηρεί σε μεγάλο βαθμό τη λογική του source dataset. Στο CIC→UNSW, τα σημαντικότερα χαρακτηριστικά παραμένουν παρόμοια με το CIC→CIC. Αυτό εξηγεί γιατί το μοντέλο αποτυγχάνει: δεν προσαρμόζεται στο UNSW-NB15, αλλά εφαρμόζει CIC-specific decision logic σε διαφορετικό περιβάλλον.
- Τρίτον, η SHAP ανάλυση επιβεβαιώνει ότι η αποτυχία δεν είναι απλώς θέμα χαμηλού accuracy. Είναι θέμα αστάθειας των εξηγήσεων. Αν ένα μοντέλο βασίζεται σε διαφορετικά χαρακτηριστικά ανάλογα με το dataset και τα χαρακτηριστικά αυτά δεν διατηρούν σταθερή σχέση με την attack κλάση, τότε δεν μπορεί να θεωρηθεί επιχειρησιακά αξιόπιστο.



Σχήμα 14: Σύγκριση SHAP Feature Importance στα Same-Dataset Πειράματα



Σχήμα 15: Σύγκριση SHAP Feature Importance στα Cross-Dataset Πειράματα

Η χρήση του SHAP ως failure analysis ενισχύει το βασικό επιχείρημα της εργασίας: η υψηλή απόδοση σε ένα μεμονωμένο dataset δεν αποδεικνύει ότι το μοντέλο έχει μάθει γενικεύσιμα μοτίβα κυβερνοεπιθέσεων. Αντίθετα, μπορεί να έχει μάθει συντομεύσεις που σχετίζονται με τη συγκεκριμένη δομή του dataset.

5.10 Σύνδεση Feature Shift και SHAP

Η ανάλυση feature distribution shift και η SHAP ανάλυση συγκλίνουν στο ίδιο συμπέρασμα. Η πρώτη δείχνει ότι τα χαρακτηριστικά έχουν διαφορετικές κατανομές μεταξύ CICIDS2017 και UNSW-NB15. Η δεύτερη δείχνει ότι το Random Forest βασίζεται σε διαφορετικά χαρακτηριστικά ανά dataset ή εφαρμόζει τη λογική του source dataset σε target dataset όπου αυτή δεν ισχύει.

Το χαρακτηριστικό dur αποτελεί χαρακτηριστικό παράδειγμα. Η τεράστια διαφορά μονάδων και τιμών μεταξύ των δύο datasets το καθιστά προβληματικό για cross-dataset μεταφορά. Παρότι ο λογαριθμικός μετασχηματισμός log_dur μειώνει την κλίμακα των τιμών, δεν εξαλείφει πλήρως τη διαφορά. Το ίδιο ισχύει και για το rate, το οποίο εμφανίζει μεγάλη διαφορά διάμεσης τιμής και IQR.

Παράλληλα, τα χαρακτηριστικά bytes και ratios, όπως sbytes, byte_ratio, pkt_ratio και bytes_per_packet, δεν έχουν την ίδια ερμηνευτική βαρύτητα στα δύο datasets. Αυτό σημαίνει ότι ακόμη και engineered features που έχουν σχεδιαστεί για να είναι πιο γενικά δεν εξασφαλίζουν από μόνα τους cross-dataset σταθερότητα.

Άρα, η αποτυχία των μοντέλων δεν πρέπει να αποδοθεί αποκλειστικά στον αλγόριθμο. Το πρόβλημα είναι συστημικό και σχετίζεται με τη μη εναρμονισμένη αναπαράσταση των δεδομένων. Τα IDS datasets διαφέρουν όχι μόνο ως προς τις επιθέσεις που περιλαμβάνουν, αλλά και ως προς τον τρόπο με τον οποίο παράγονται τα χαρακτηριστικά.

5.11 Επιχειρησιακές Επιπτώσεις

Τα αποτελέσματα έχουν άμεση σημασία για πραγματικά περιβάλλοντα κυβερνοασφάλειας. Ένα μοντέλο IDS που εκπαιδεύεται σε δημόσιο benchmark και εμφανίζει υψηλό F1-score δεν είναι απαραίτητα κατάλληλο για χρήση σε διαφορετικό δίκτυο. Αν το δίκτυο παραγωγής έχει διαφορετική κατανομή χαρακτηριστικών από το dataset εκπαίδευσης, τότε το μοντέλο μπορεί να αποτύχει με απρόβλεπτο τρόπο.

Η αποτυχία CIC→UNSW αντιστοιχεί σε ένα IDS που δεν ανιχνεύει καμία επίθεση. Αυτό είναι ίσως το πιο επικίνδυνο σενάριο, επειδή το σύστημα φαίνεται ήσυχο, αλλά στην πραγματικότητα δεν προστατεύει το δίκτυο.

Η αποτυχία UNSW→CIC αντιστοιχεί σε ένα IDS που παράγει υπερβολικά πολλούς συναγερμούς. Αυτό δεν είναι ασφαλές ούτε πρακτικά χρήσιμο, επειδή οι αναλυτές ασφαλείας δεν μπορούν να διαχειριστούν τόσο μεγάλο όγκο false positives.

Επομένως, πριν από την επιχειρησιακή χρήση ML-based IDS απαιτούνται:

1. εναρμόνιση των χαρακτηριστικών,
2. έλεγχος μονάδων μέτρησης,
3. αξιολόγηση σε περισσότερα από ένα datasets,
4. cross-dataset testing,
5. ανάλυση feature shift,
6. ερμηνευσιμότητα μέσω XAI,
7. αξιολόγηση όχι μόνο με accuracy αλλά και με recall, precision, F1 και balanced accuracy.

Χωρίς αυτά τα βήματα, η απόδοση ενός IDS σε benchmark μπορεί να είναι παραπλανητική.

5.12 Περιορισμοί της Ανάλυσης

Παρότι η ανάλυση παρέχει σημαντικές ενδείξεις για τα αίτια της αποτυχίας γενίκευσης, υπάρχουν ορισμένοι περιορισμοί.

- Πρώτον, η ανάλυση βασίζεται σε δύο datasets, το CICIDS2017 και το UNSW-NB15. Αν και πρόκειται για δημοφιλή benchmarks, η χρήση περισσότερων datasets θα επέτρεπε πιο γενικευμένα συμπεράσματα.
- Δεύτερον, η εργασία χρησιμοποιεί κοινό feature space 11 χαρακτηριστικών. Η επιλογή αυτή είναι απαραίτητη για τη cross-dataset αξιολόγηση, αλλά αποκλείει άλλα χαρακτηριστικά που ίσως είναι χρήσιμα σε κάθε dataset ξεχωριστά.
- Τρίτον, η εναρμόνιση των χαρακτηριστικών είναι σημασιολογική και όχι πλήρως στατιστική. Η εργασία αναδεικνύει τη σημασία της διαφοράς μονάδων και κατανομών, αλλά δεν εφαρμόζει πλήρες domain adaptation ή πλήρη επανεξαγωγή χαρακτηριστικών από raw PCAP αρχεία.
- Τέταρτον, η SHAP ανάλυση εφαρμόστηκε μόνο στο Random Forest. Η επιλογή αυτή είναι μεθοδολογικά δικαιολογημένη, καθώς το Random Forest ήταν το ισχυρότερο μοντέλο στα same-dataset σενάρια, αλλά δεν επιτρέπει πλήρη σύγκριση XAI εξηγήσεων μεταξύ όλων των μοντέλων.
- Πέμπτον, το SHAP εξηγεί τη συμπεριφορά του μοντέλου, όχι την αντικειμενική αιτία της επίθεσης. Επομένως, οι SHAP τιμές πρέπει να ερμηνεύονται ως εξηγήσεις των αποφάσεων του μοντέλου και όχι ως απόλυτη απόδειξη αιτιότητας.

5.13 Σύνοψη Κεφαλαίου

Στο παρόν κεφάλαιο αναλύθηκαν οι αιτίες της αποτυχίας των μοντέλων στα cross-dataset πειράματα. Η ανάλυση feature distribution shift έδειξε ότι τα χαρακτηριστικά του CICIDS2017 και του UNSW-NB15 παρουσιάζουν σημαντικές στατιστικές διαφορές, ακόμη και όταν έχουν παρόμοια σημασιολογική έννοια. Ιδιαίτερα χαρακτηριστικά όπως το dur, το rate, το sbytes και το byte_ratio εμφανίζουν μεγάλες αποκλίσεις σε διάμεσο και IQR.

Η ανάλυση του RobustScaler έδειξε ότι η κανονικοποίηση, παρότι απαραίτητη και σωστή για την αποφυγή data leakage, δεν αρκεί για να λύσει το πρόβλημα της μετατόπισης

κατανομής. Όταν ο scaler εκπαιδεύεται σε ένα dataset και εφαρμόζεται σε ένα άλλο, οι τιμές του target dataset τοποθετούνται σε feature space που έχει οριστεί από το source dataset.

Η SHAP ανάλυση του Random Forest έδειξε ότι τα χαρακτηριστικά που επηρεάζουν τις αποφάσεις του μοντέλου δεν παραμένουν σταθερά μεταξύ datasets. Στο CICIDS2017 κυριαρχούν byte-related χαρακτηριστικά όπως bytes_per_packet και sbytes, ενώ στο UNSW-NB15 μεγαλύτερη σημασία έχουν αναλογικά και rate-based χαρακτηριστικά όπως byte_ratio, pkt_ratio, rate και log_rate.

Συνολικά, το κεφάλαιο κατέδειξε ότι η αποτυχία cross-dataset γενίκευσης δεν είναι απλώς αδυναμία ενός συγκεκριμένου αλγορίθμου. Είναι αποτέλεσμα βαθύτερης ασυμβατότητας μεταξύ datasets, διαφορετικών κατανομών χαρακτηριστικών, διαφορετικών μονάδων μέτρησης και dataset-specific μοτίβων. Το συμπέρασμα αυτό ενισχύει το κεντρικό επιχείρημα της εργασίας: η αξιολόγηση ML-based IDS πρέπει να ξεπερνά τα same-dataset benchmarks και να περιλαμβάνει cross-dataset, feature-shift και explainability analysis.

6. Συμπεράσματα και Μελλοντική Έρευνα

Η παρούσα διπλωματική εργασία εξέτασε το πρόβλημα της γενίκευσης μοντέλων Μηχανικής Μάθησης σε συστήματα ανίχνευσης εισβολών δικτύου. Το βασικό ερευνητικό ερώτημα ήταν αν ένα μοντέλο που εκπαιδεύεται σε ένα δημόσια διαθέσιμο dataset μπορεί να διατηρήσει ικανοποιητική απόδοση όταν αξιολογείται σε ένα διαφορετικό dataset, με διαφορετική κατανομή χαρακτηριστικών, διαφορετική μεθοδολογία συλλογής δεδομένων και διαφορετικό περιβάλλον δικτυακής κίνησης.

Για την απάντηση του ερωτήματος χρησιμοποιήθηκαν δύο γνωστά datasets στον χώρο των Network Intrusion Detection Systems, το CICIDS2017 και το UNSW-NB15. Η πειραματική διαδικασία βασίστηκε σε κοινό feature space 11 χαρακτηριστικών, σε δυαδική ταξινόμηση benign/attack και σε τρία supervised Machine Learning μοντέλα: Logistic Regression, Random Forest και LinearSVC.

Τα μοντέλα αξιολογήθηκαν σε τέσσερα σενάρια: δύο same-dataset πειράματα, CIC→CIC και UNSW→UNSW, και δύο cross-dataset πειράματα, CIC→UNSW και UNSW→CIC. Επιπλέον, εφαρμόστηκε ανάλυση μετατόπισης κατανομής χαρακτηριστικών και SHAP-based explainable failure analysis για την ερμηνεία της συμπεριφοράς του Random Forest.

6.1 Βασικά Συμπεράσματα της Εργασίας

Το πρώτο βασικό συμπέρασμα είναι ότι τα μοντέλα Μηχανικής Μάθησης μπορούν να επιτύχουν πολύ υψηλή απόδοση όταν εκπαιδεύονται και αξιολογούνται στο ίδιο dataset. Ειδικά το Random Forest παρουσίασε εξαιρετικά αποτελέσματα στα same-dataset πειράματα, με F1-score 0,952 στο CIC→CIC και 0,931 στο UNSW→UNSW. Τα αποτελέσματα αυτά δείχνουν ότι το μοντέλο μπορεί να μάθει αποτελεσματικά τα μοτίβα που υπάρχουν στο dataset εκπαίδευσης.

Ωστόσο, το δεύτερο και σημαντικότερο συμπέρασμα είναι ότι η υψηλή same-dataset απόδοση δεν συνεπάγεται πραγματική ικανότητα γενίκευσης. Όταν τα ίδια μοντέλα αξιολογήθηκαν σε διαφορετικό dataset από αυτό στο οποίο εκπαιδεύτηκαν, η απόδοση μειώθηκε δραματικά. Στο CIC→UNSW, το Random Forest είχε F1-score 0,000, καθώς δεν εντόπισε καμία επίθεση. Στο UNSW→CIC, το ίδιο μοντέλο είχε F1-score μόλις 0,172, με πολύ χαμηλό precision λόγω του τεράστιου αριθμού false positives.

Το τρίτο συμπέρασμα είναι ότι η αποτυχία γενίκευσης δεν εμφανίζεται πάντα με τον ίδιο τρόπο. Στο CIC→UNSW παρατηρήθηκε σιωπηλή αποτυχία, καθώς το μοντέλο χαρακτήρισε όλες τις εγγραφές ως benign και δεν ανίχνευσε καμία κακόβουλη δραστηριότητα. Αντίθετα, στο UNSW→CIC παρατηρήθηκε θορυβώδης αποτυχία, καθώς το μοντέλο εντόπισε αρκετές επιθέσεις, αλλά παρήγαγε υπερβολικά μεγάλο αριθμό ψευδών συναγερμών.

Το τέταρτο συμπέρασμα είναι ότι η χρήση accuracy ως κύριας μετρικής μπορεί να είναι παραπλανητική σε IDS datasets. Σε περιπτώσεις έντονης ανισορροπίας κλάσεων, ένα μοντέλο μπορεί να εμφανίζει υψηλό accuracy χωρίς να έχει ουσιαστική ικανότητα ανίχνευσης επιθέσεων. Για τον λόγο αυτό, οι μετρικές F1-score, recall, precision, balanced accuracy και confusion matrices είναι καταλληλότερες για την αξιολόγηση συστημάτων ανίχνευσης εισβολών.

6.2 Ερμηνεία της Αποτυχίας Γενίκευσης

Η αποτυχία των μοντέλων στα cross-dataset πειράματα δεν μπορεί να αποδοθεί αποκλειστικά στην επιλογή αλγορίθμου. Και τα τρία μοντέλα απέτυχαν να γενικεύσουν ικανοποιητικά, παρότι ανήκουν σε διαφορετικές κατηγορίες ταξινομητών. Αυτό δείχνει

ότι το πρόβλημα είναι βαθύτερο και σχετίζεται με τη φύση των datasets και του feature space.

Η ανάλυση μετατόπισης κατανομής χαρακτηριστικών έδειξε ότι τα ίδια σημασιολογικά χαρακτηριστικά έχουν διαφορετική στατιστική συμπεριφορά στα δύο datasets. Χαρακτηριστικά όπως το dur, το rate, το sbytes, το byte_ratio και το pkt_ratio εμφανίζουν σημαντικές αποκλίσεις σε διάμεσο και ενδοτεταρτημοριακό εύρος. Ιδιαίτερα το dur αποτελεί χαρακτηριστικό παράδειγμα προβληματικής αντιστοίχισης, καθώς στο CICIDS2017 εκφράζεται σε μικροδευτερόλεπτα, ενώ στο UNSW-NB15 σε δευτερόλεπτα.

Η χρήση RobustScaler ήταν μεθοδολογικά σωστή για την αποφυγή data leakage, καθώς ο scaler εκπαιδεύτηκε μόνο στα δεδομένα εκπαίδευσης. Ωστόσο, η κανονικοποίηση δεν μπορεί να λύσει από μόνη της το πρόβλημα της διαφορετικής κατανομής χαρακτηριστικών. Όταν ο scaler μαθαίνει τα στατιστικά του source dataset και εφαρμόζεται στο target dataset, οι τιμές του target dataset τοποθετούνται σε feature space που έχει οριστεί από το source dataset. Αυτό μπορεί να οδηγήσει σε πλήρως λανθασμένη συμπεριφορά του μοντέλου.

Η SHAP ανάλυση ενίσχυσε αυτή την ερμηνεία. Στο CICIDS2017 το Random Forest βασίστηκε κυρίως σε byte-related χαρακτηριστικά, όπως bytes_per_packet, sbytes και byte_ratio. Στο UNSW-NB15, μεγαλύτερη σημασία είχαν χαρακτηριστικά όπως byte_ratio, pkt_ratio, rate και log_rate. Αυτό δείχνει ότι το μοντέλο δεν χρησιμοποιεί τα ίδια χαρακτηριστικά με τον ίδιο τρόπο στα δύο datasets. Επομένως, η υψηλή απόδοση σε ένα dataset μπορεί να οφείλεται σε dataset-specific patterns και όχι σε γενικεύσιμη κατανόηση της κακόβουλης δικτυακής συμπεριφοράς.

6.3 Επιστημονική και Πρακτική Συμβολή

Η παρούσα εργασία συμβάλλει στη μελέτη των ML-based IDS με τέσσερις βασικούς τρόπους.

- Πρώτον, αναδεικνύει πειραματικά ότι η αξιολόγηση μοντέλων μόνο σε train/test split του ίδιου dataset δεν είναι επαρκής. Ένα μοντέλο μπορεί να εμφανίζεται εξαιρετικά αποτελεσματικό σε same-dataset συνθήκες, αλλά να αποτυγχάνει πλήρως όταν εφαρμόζεται σε διαφορετικό dataset.
- Δεύτερον, δείχνει ότι η cross-dataset αξιολόγηση είναι απαραίτητη για την εκτίμηση της πραγματικής αξιοπιστίας ενός IDS. Η επιχειρησιακή χρήση ενός μοντέλου σε

πραγματικό δίκτυο δεν μοιάζει με αξιολόγηση σε ελεγχόμενο benchmark. Τα πραγματικά δίκτυα έχουν διαφορετικές συσκευές, διαφορετικά πρωτόκολλα, διαφορετική ένταση κίνησης και διαφορετικά μοτίβα επιθέσεων.

- Τρίτον, η εργασία αναδεικνύει τη σημασία της εναρμόνισης του feature space. Η απλή αντιστοίχιση ονομάτων χαρακτηριστικών δεν αρκεί. Χρειάζεται έλεγχος μονάδων μέτρησης, στατιστική ανάλυση κατανομών και ενδεχομένως επανεξαγωγή χαρακτηριστικών από raw δεδομένα με ενιαία διαδικασία.
- Τέταρτον, η χρήση SHAP ως explainable failure analysis προσφέρει μια πιο ουσιαστική κατανόηση της συμπεριφοράς του μοντέλου. Η ερμηνευσιμότητα δεν χρησιμοποιείται απλώς για να εξηγήσει σωστές προβλέψεις, αλλά για να αποκαλύψει γιατί ένα μοντέλο που φαίνεται ισχυρό σε ένα dataset αποτυγχάνει σε ένα άλλο.

6.4 Περιορισμοί της Εργασίας

Παρότι η εργασία προσφέρει χρήσιμα συμπεράσματα, υπάρχουν ορισμένοι περιορισμοί που πρέπει να αναφερθούν.

- Πρώτον, χρησιμοποιήθηκαν μόνο δύο datasets, το CICIDS2017 και το UNSW-NB15. Αν και πρόκειται για δύο από τα πιο διαδεδομένα datasets στον χώρο των IDS, η χρήση περισσότερων datasets, όπως IoT-23, BoT-IoT ή CICIoT2023, θα επέτρεπε ισχυρότερα και πιο γενικεύσιμα συμπεράσματα.
- Δεύτερον, η εργασία περιορίστηκε σε δυαδική ταξινόμηση benign/attack. Η επιλογή αυτή ήταν απαραίτητη για τη συγκρισιμότητα μεταξύ datasets, αλλά δεν επιτρέπει ανάλυση ανά τύπο επίθεσης. Σε πραγματικά περιβάλλοντα, η διάκριση μεταξύ διαφορετικών τύπων επιθέσεων μπορεί να είναι εξίσου σημαντική με τη γενική διάκριση benign/attack.
- Τρίτον, χρησιμοποιήθηκε περιορισμένο κοινό feature space 11 χαρακτηριστικών. Αυτό επέτρεψε τη cross-dataset αξιολόγηση, αλλά απέκλεισε άλλα χαρακτηριστικά που ενδέχεται να είναι χρήσιμα εντός κάθε dataset ξεχωριστά.
- Τέταρτον, δεν αξιολογήθηκαν μοντέλα Deep Learning. Η απόφαση αυτή ήταν συνειδητή, καθώς ο στόχος της εργασίας ήταν η καθαρή ανάλυση της γενίκευσης κλασικών supervised ML μοντέλων. Παρόλα αυτά, η σύγκριση με DL αρχιτεκτονικές θα μπορούσε να αποτελέσει μελλοντική επέκταση.
- Πέμπτον, η SHAP ανάλυση εφαρμόστηκε μόνο στο Random Forest. Η επιλογή αυτή δικαιολογείται από το γεγονός ότι το Random Forest ήταν το ισχυρότερο μοντέλο στα

same-dataset πειράματα, αλλά δεν επιτρέπει πλήρη σύγκριση εξηγήσεων μεταξύ όλων των μοντέλων.

- Έκτον, δεν εφαρμόστηκε πλήρης domain adaptation ή πλήρης εναρμόνιση των μονάδων μέτρησης. Αντίθετα, η εργασία χρησιμοποίησε τη μη εναρμονισμένη μεταφορά ως μέσο ανάδειξης του προβλήματος feature distribution shift. Αυτό είναι χρήσιμο ερευνητικά, αλλά δεν αποτελεί τελική λύση για παραγωγικό IDS.

6.5 Προτάσεις για Μελλοντική Έρευνα

Η παρούσα εργασία ανοίγει αρκετές κατευθύνσεις για μελλοντική έρευνα.

- Μία πρώτη κατεύθυνση είναι η πλήρης εναρμόνιση των χαρακτηριστικών πριν από την εκπαίδευση και αξιολόγηση. Αυτό περιλαμβάνει μετατροπή μονάδων μέτρησης, όπως η μετατροπή της διάρκειας ροής σε κοινή μονάδα, καθώς και στατιστική κανονικοποίηση χαρακτηριστικών με τρόπο που να είναι κατάλληλος για cross-dataset μεταφορά.
- Μία δεύτερη κατεύθυνση είναι η επανεξαγωγή χαρακτηριστικών από raw PCAP αρχεία με ενιαίο εργαλείο και ενιαία διαδικασία. Αν τα datasets παράγονται με διαφορετικά flow exporters, τότε ακόμη και χαρακτηριστικά με ίδιο όνομα μπορεί να μην είναι πλήρως συγκρίσιμα. Η χρήση ενιαίου εργαλείου εξαγωγής θα μπορούσε να μειώσει σημαντικά την ασυμβατότητα του feature space.
- Μία τρίτη κατεύθυνση είναι η αξιολόγηση σε περισσότερα datasets. Η προσθήκη datasets όπως IoT-23, BoT-IoT, CICIoT2023 και ToN-IoT θα επέτρεπε την εξέταση της γενίκευσης σε πιο σύνθετα και ετερογενή περιβάλλοντα, ιδιαίτερα σε IoT και βιομηχανικά δίκτυα.
- Μία τέταρτη κατεύθυνση είναι η χρήση τεχνικών domain adaptation και transfer learning. Τέτοιες τεχνικές θα μπορούσαν να βοηθήσουν τα μοντέλα να προσαρμόζονται σε νέο target dataset χωρίς να απαιτείται πλήρης επανεκπαίδευση από την αρχή.
- Μία πέμπτη κατεύθυνση είναι η ενσωμάτωση online learning ή incremental learning. Σε πραγματικά δίκτυα, η κατανομή της κίνησης αλλάζει με τον χρόνο. Επομένως, ένα IDS που μπορεί να ενημερώνεται σταδιακά με νέα δεδομένα θα ήταν πιο κατάλληλο για επιχειρησιακή χρήση.
- Μία έκτη κατεύθυνση είναι η επέκταση της explainability analysis. Μελλοντικές εργασίες μπορούν να συγκρίνουν SHAP, LIME, Anchors και counterfactual

explanations, όχι μόνο ως εργαλεία ερμηνείας, αλλά και ως μηχανισμούς ελέγχου της σταθερότητας των χαρακτηριστικών μεταξύ datasets.

- Μία έβδομη κατεύθυνση είναι η πολυταξική ταξινόμηση επιθέσεων. Αντί για απλή διάκριση benign/attack, θα μπορούσε να εξεταστεί η ικανότητα των μοντέλων να αναγνωρίζουν συγκεκριμένους τύπους επιθέσεων, όπως DoS, DDoS, brute force, reconnaissance, exploits και botnet traffic.
- Τέλος, σημαντική κατεύθυνση αποτελεί η αξιολόγηση robustness απέναντι σε adversarial attacks. Τα ML-based IDS μπορεί να είναι ευάλωτα σε μικρές μεταβολές των χαρακτηριστικών εισόδου. Επομένως, η μελλοντική έρευνα πρέπει να εξετάσει όχι μόνο αν τα μοντέλα γενικεύουν, αλλά και αν παραμένουν ανθεκτικά απέναντι σε σκόπιμη παραπλάνηση.

6.6 Τελικό Συμπέρασμα

Το τελικό συμπέρασμα της εργασίας είναι σαφές: η υψηλή απόδοση ενός Machine Learning μοντέλου σε ένα μεμονωμένο IDS dataset δεν αποτελεί επαρκή ένδειξη πραγματικής επιχειρησιακής αξιοπιστίας. Τα same-dataset πειράματα μπορούν να υπερεκτιμήσουν σημαντικά την ικανότητα ενός μοντέλου, επειδή το train και το test set προέρχονται από την ίδια κατανομή.

Η cross-dataset αξιολόγηση έδειξε ότι τα μοντέλα που φαίνονται ισχυρά σε ένα dataset μπορούν να αποτύχουν πλήρως σε ένα άλλο. Η αποτυχία αυτή δεν είναι απλώς αποτέλεσμα κακής επιλογής αλγορίθμου. Συνδέεται με βαθύτερα προβλήματα, όπως η διαφορετική διαδικασία συλλογής δεδομένων, η ασυμβατότητα χαρακτηριστικών, η διαφορά μονάδων μέτρησης, η ανισορροπία κλάσεων και η μετατόπιση κατανομής.

Η εργασία δείχνει ότι η ανάπτυξη αξιόπιστων ML-based IDS απαιτεί περισσότερα από υψηλό accuracy σε ένα benchmark. Απαιτεί ελεγχόμενη προεπεξεργασία, κοινό και εναρμονισμένο feature space, cross-dataset αξιολόγηση, κατάλληλες μετρικές και ερμηνευσιμότητα των αποφάσεων. Χωρίς αυτά, ένα μοντέλο μπορεί να φαίνεται αποτελεσματικό σε εργαστηριακό περιβάλλον, αλλά να είναι αναξιόπιστο ή ακόμη και επικίνδυνο σε πραγματικές συνθήκες.

Συνεπώς, η βασική συμβολή της παρούσας διπλωματικής είναι ότι αναδεικνύει την ανάγκη μετάβασης από απλή same-dataset αξιολόγηση σε πιο αυστηρά, εξηγήσιμα και

ρεαλιστικά cross-dataset πειραματικά πλαίσια για την ανάπτυξη συστημάτων ανίχνευσης εισβολών με Μηχανική Μάθηση.

7. Βιβλιογραφία

1. Ahmed Abd El-Baset Donkol, A. G. (2018). *Optimization of Intrusion Detection Using Likely Point PSO and Enhanced LSTM-RNN Hybrid Technique in Communication Networks*. Retrieved from IEEE Xplore: <https://ieeexplore.ieee.org/document/10026342?denied=>
2. Alaa Alhowaide, I. A. (2022). *Balanced Datasets for IoT IDS*. Retrieved from Cornell University: <https://arxiv.org/abs/2301.04008>
3. Arnaud Rosay, E. C. (2022). *Network Intrusion Detection: A Comprehensive Analysis of CIC-IDS2017*. Retrieved from HAL: <https://hal.science/hal-03563228/>
4. Danish Javeed, T. G. (2024). *An Explainable and Resilient Intrusion*. Retrieved from IEEE Xplore: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10146000>
5. Emanuel Krzysztoń, I. R. (2024). *A Comparative Analysis of Anomaly Detection Methods in IoT Networks: An Experimental Study*. Retrieved from MDPI: <https://www.mdpi.com/2076-3417/14/24/11545>
6. Euclides Carlos Pinto Neto, S. D. (2023). *CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment*. Retrieved from MDPI: <https://www.mdpi.com/1424-8220/23/13/5941>
7. G., M. (2023). *E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection*. Retrieved from Peireas University.
8. Makridis, G. (2024). *An Explainable and Resilient Intrusion Detection System for Industry 5.0*. Retrieved from Peireas University.
9. Maria Carla Galzarossa, P. G. (2025). An assessment framework for explainable AI with applications to cybersecurity. *Artificial Intelligence Review*. Retrieved from Artificial Intelligence Review.
10. María Rodríguez, Á. A. (2022). *Evaluation of Machine Learning Techniques for Traffic Flow-Based Intrusion Detection*. Retrieved from MDPI: <https://www.mdpi.com/1424-8220/22/23/9326>

11. Marta Catillo, A. D. (2023). *A Case Study with CICIDS2017 on the Robustness of Machine Learning against Adversarial Attacks in Intrusion Detection*. Retrieved from ACM DL DIGITAL LIBRARY: <https://dl.acm.org/doi/10.1145/3600160.3605031>
12. Mohanad Sarhan, S. L. (2024). *Feature extraction for machine learning-based intrusion detection in IoT networks*. Retrieved from Science Direct: <https://www.sciencedirect.com/science/article/pii/S2352864822001754>
13. Noveela Iftikhar, M. U. (2025). *Intrusion Detection in NSL-KDD Dataset Using Hybrid Self-Organizing Map Model*. Retrieved from Tech Science Press: <https://www.techscience.com/CMES/v143n1/60475/html>
14. Nuno Prazeres. (2023). *Engineering the application of machine learning in an IDS based on IoT traffic flow*. Retrieved from Science Direct: <https://www.sciencedirect.com/science/article/pii/S2667305323000145>
15. Osvaldo Arreche, T. R. (2023). *E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection*. Retrieved from IEEE Xplore: <https://ieeexplore.ieee.org/document/10433134?denied=>
16. Rudresh, D., Devam, D., Het, N., Smiti, S., Rana, O., Pankesh , P., . . . Rajiv, R. (2023). *Explainable AI (XAI): Core Ideas, Techniques and Solutions*. Retrieved from Cardiff University: <https://dl.acm.org/doi/abs/10.1145/3561048>
17. Shweta More, M. I. (2024). *Enhanced Intrusion Detection Systems Performance with UNSW-NB15 Data Analysis*. Retrieved from MDPI: <https://www.mdpi.com/1999-4893/17/2/64>
18. Singh Deepika, M. S. (2024). Identifying Cyber Threats in Metaverse Learning Environment using Explainable Deep Neural Networks. *International Journal of Performability Engineering*.
19. Sunday Adeola Ajagbe, J. B. (2024). *Intrusion Detection: A Comparison Study of Machine Learning Models Using Unbalanced Dataset*. Retrieved from Springer Nature Link: <https://link.springer.com/article/10.1007/s42979-024-03369-0>
20. Vibhute, A. D. (2024). *Network anomaly detection and performance evaluation of Convolutional Neural Networks on UNSW-NB15 dataset*. Retrieved from SCIENCE DIRECT: <https://www.sciencedirect.com/science/article/pii/S1877050924008871>
21. Wali Syed, F. Y. (2024). *Meta: Toward a Unified, Multimodal Dataset for Network Intrusion Detection Systems*. Retrieved from IEEE Xplore: <https://ieeexplore.ieee.org/abstract/document/10720901>
22. Wali Syed, F. Y. (2025). Explainable AI and Random Forest based reliable intrusion detection system. *Computers & Security*.

23. Yang, L. (2023). *Enabling AutoML for Zero-Touch Network Security: Use-Case*. Retrieved from Google Scholar:
https://d197for5662m48.cloudfront.net/documents/publicationstatus/171716/preprint_pdf/c3786e7dbc9cb693586f5b0562da67c5.pdf